Lecture Notes in
Computer Science 1818

Cambyse Guy Omidyar (Ed.)

Mobile and Wireless Communications Networks

IFIP-TC6/European Commission NETWORKING 2000
International Workshop, MWCN 2000
Paris, France, May 2000
Proceedings

IFIP TC6

Springer

Lecture Notes in Computer Science 1818
Edited by G. Goos, J. Hartmanis and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Cambyse Guy Omidyar (Ed.)

Mobile and Wireless Communications Networks

IFIP-TC6/European Commission NETWORKING 2000
International Workshop, MWCN 2000
Paris, France, May 16-17, 2000
Proceedings

Springer

Volume Editor

Cambyse Guy Omidyar
Computer Sciences Corporation
7459-A Candlewood Drive
Hanover, Maryland 21076, USA
E-mail: gomidyar@csc.com

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Mobile and wireless communication networks : European union networking 2000 ; international workshop ; proceedings / IFIP TC 6, MWCN 2000, Paris, France, May 16 - 17, 2000. Cambyse Guy Omidyar (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Singapore ; Tokyo : Springer, 2000
(Lecture notes in computer science ; Vol. 1818)
ISBN 3-540-67543-4

CR Subject Classification (1998): C.2, C.4, D.2, H.4.3, J.2, J.1

ISSN 0302-9743
ISBN 3-540-67543-4 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag is a company in the BertelsmannSpringer publishing group.

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign
Printed on acid-free paper SPIN 10720335 06/3142 5 4 3 2 1 0

Preface

o l l o u t o N two k (N w th
o o t t t ul t g t h l h g th
g g o t t fi l o o l l o u t o N two k
h fi t o took l t So o v ty y 1 1
199

t g t o o fi o t l w l t h olog th o l ty
tu to u t fi o l two k t o t ff -
t v ffi t w y to ov l -to- o t v ty u u tou
k t wh o w l two k h v g ow ly
wh h t to g t ll o o oll v u h loy t o
o ll- t h olog lu g o t t o l ty
th t g t o w th g g o w l two k (N
o g gly o t t o l ty o t o g t lo-
t o g t o t o t l h t u l ty o v ov o g
t -wo k g o w -l w th w l two k o vo v o g
t w o o th o o u w l two k g o th o
th l t w y w th th v lo t o g t l t h olog t ll t h v
g th o o t o ult l v y t h u o u g u t
g t l o t y t o w to v lo g o h t t l t o
o t g t v t ll t two k g ow w ll- t l h olut o
o th o o th w ll u h wo k ho w t to
ov t ly o u o lo to y h t l o t ut o o
No th u o h l t h t t h l g
to t y th th o t l tu l o l th t to y o l o -
l o t w th o l w l o u t o l k two k
h o w o o y 6 u o o u N -
t o l l h h l o u t o (N o-o t o w th
S L th o So h l o tt o o u -
t o Sy t t g t o o l g took l t th t S
L ll tt y 16 1

h uy y

Scientific Program Committee

ghv g g
v t
h
g v ty S
y v ty gy t t
tt t t y
S
h v S
v
u t
g v ty S
z v t ut u
t
- t t t S
gh h h t h S
y S S h St g tt h
h v t yt h t tut S
S S
h tz h g h gt v ty S
u v t h
g v ty
u S
u Sw tz
S Sh ug v ty S
S hw tz u v ty S
S v t z h
S h v ty h y
S S u
gh h

Table of Contents

Indoor Wireless Networking

Multiple Access Techniques for Wireless Ad-Hoc Networking

Chair: Khaled Elsayed, Cairo University, Egypt

Telephony over Packet Switched Networks

Chair: Prathima Agrawal, Telecordia, USA

IP Networks Versus Conventional Switched Networks

Chair: Pascal Lorenz, Université d'Haute Alsace, France

Mobility Management and Access Techniques

Chair: Mahmoud Naghshineh, IBM,
Thomas J. Watson Research Center, USA

Mobility Support in IP

Chair: Kaveh Pahlavan, Worcester Polytechnic
Institute, USA

An Overview of Wireless Indoor Geolocation Techniques and Systems

Kaveh Pahlavan, Xinrong Li, Mika Ylianttila, Ranvir Chana, and Matti Latva-aho

Center for Wireless Information Network Studies, Worcester Polytechnic Institute, USA
{kaveh, xinrong}@ece.wpi.edu
Centre for Wireless Communications, University of Oulu, Finland
{over, rschana, matla}@ees2.oulu.fi

Abstract. Wireless indoor networks are finding their way into the home and office environments. Also, exploiting location information becomes very popular for both wireless service providers and consumers applications. However, the indoor radio channel causes challenges in extracting accurate location information in indoor environment so that traditional GPS and cellular location systems cannot work properly in indoor areas. This paper provides an overview of the indoor geolocation techniques. After introducing an overall architecture for indoor geolocation systems, technical overview of two indoor geolocation systems are presented. To demonstrate the predicted performance of such systems some simulation results obtained from an indoor geolocation demonstrator are presented.

1. Introduction

Today numerous wireless indoor products such as cordless telephone, wireless security systems, cordless speakers and even wireless Internet access have been introduced to the consumers. The same way as in late 70's and early 80's the increasing number of terminals in the offices initiated the LAN industry, today increasing number of wireless terminals in indoor areas is promoting wireless indoor networking. Indoor areas are difficult for wiring and most people are reluctant to allow workers to do extensive wiring inside buildings. Wireless is a cost efficient solution that can also provide additional feature of mobility and convenient relocatability. These reasons leave wireless as the preferred medium for indoor networking in the future and promote local networking activities such as Bluetooth, Home-RF, IEEE 802.11 and HIPERLAN/2.

An important evolving technology in recent years has been the indoor geolocation technology both for military and commercial applications. In the commercial application there is an increasing need for these systems for application in hospitals to locate patients or expensive equipment and in homes to locate children and equipment. Military and public safety applications in urban scenarios have promoted a need for inbuilding communication and geolocation networks enabling soldiers, policeman, and fire fighters to complete their missions in urban areas. These incentives have lead to research in indoor geolocation systems [1][2][3]. Due to indoor path loss, traditional GPS or E-911 location system cannot work properly in

C.G. Omidyar (Ed.): MWCN 2000, LNCS 1818, pp. 1-13, 2000.

indoor environment. As a result, dedicated indoor geolocation systems have to be developed to provide accurate indoor geolocation services.

This paper provides an update on the trends in indoor geolocation systems. In section 2, we briefly discuss the overall system architecture and various geolocation metrics that can be used in indoor environment. Then technical overviews of two indoor geolocation products are presented in Section 3. In Section 4, some simulation results obtained from an indoor geolocation testbed are included. Finally, we close this paper with a short conclusion.

2. Wireless Geolocation Methods and Metrics

Most of the geolocation system architectures and methods developed for cellular systems are applicable for indoor geolocation systems although special considerations are needed for indoor radio channels. The most widely used wireless geolocation metrics include Angel of Arrival (AOA), Time of Arrival (TOA), Time Differences of Arrival (TDOA), Received Signal Strength (RSS) and Received Signal Phase. In this section we present an overview of overall system architectures and basic concepts of geolocation metrics as well as corresponding geolocation methods. The possibility of using these methods in indoor environment is also considered.

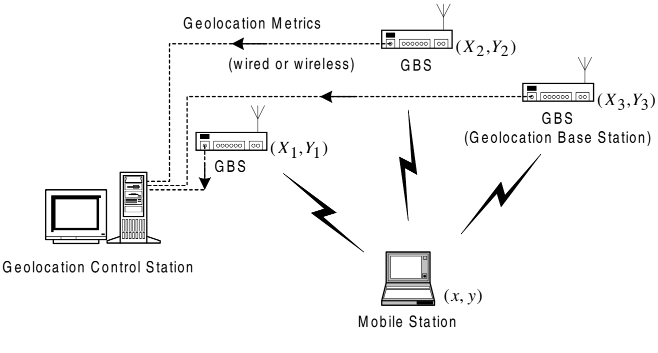


Fig. 1. Overall architecture of indoor geolocation system.

2.1 Overall System Architecture

Similar to the cellular geolocation system, the architecture of indoor geolocation systems can be roughly grouped into two main categories: mobile-based architecture and network-based architecture. Most of the indoor geolocation applications proposed to date have been focused on network-based system architecture as shown in Fig. 1 [4][5]. The geolocation base stations (GBS) extract location metrics from the radio signals transmitted by the mobile station and relay the information to a geolocation control station (GCS). The connection between GBS and GCS can be either wired or wireless. Then the position of the mobile station is estimated, displayed and tracked at the GCS. With the mobile-based system architecture, the mobile station estimates self-position by measuring received radio signals from multiple fixed GBS. Compared to mobile-based architecture, the network-based system has the advantage that the mobile station can be implemented as a simple-structured transceiver with small size and low power consumption that can be easily carried by people or attached to valuable equipments as a tag.

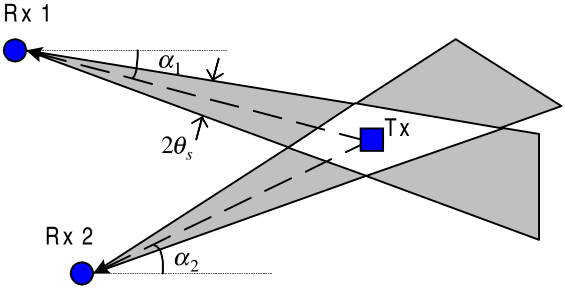


Fig. 2. Angel of Arrival geolocation method.

2.2 Angle of Arrival

The AOA geolocation method uses simple triangulation to locate the transmitter as shown in Fig. 2. The receiver measures the direction of received signals (i.e. angel of arrival) from the target transmitter using directional antennas or antenna arrays. If the accuracy of the direction measurement is $\pm\theta_s$, AOA measurement at the receiver will restrict the transmitter position around the line-of-sight (LOS) signal path with an angular spread of $2\theta_s$. AOA measurements at two receivers will provide a position fix as illustrated in Fig. 2. We can clearly observe that given the accuracy of AOA measurement, the accuracy of the position estimation depends on the transmitter position with respect to the receivers. When the transmitter lies between the two receivers, AOA measurements will not be able to provide a position fix. As a result, more than two receivers are normally needed to improve the location accuracy. For macro-cellular environment where the primary scatters are located around the transmitter and far away from the receivers, AOA method can provide acceptable location accuracy [6]. But dramatically large location errors will occur if the LOS signal path is blocked and the AOA of a reflected or a scattered signal component is used for estimation. In indoor environment, the LOS signal path is usually blocked by surrounding objects or walls. Thus AOA method will not be usable as the only metric for indoor geolocation system.

2.3 Time of Arrival and Time Difference of Arrival

The TOA method is based on estimating the propagation time of the signals (i.e. TOA) from a transmitter to multiple receivers. Several different methods can be used to obtain TOA or TDOA estimates, including pulse ranging [9][10], phase ranging [9] and spread-spectrum techniques [4][11].

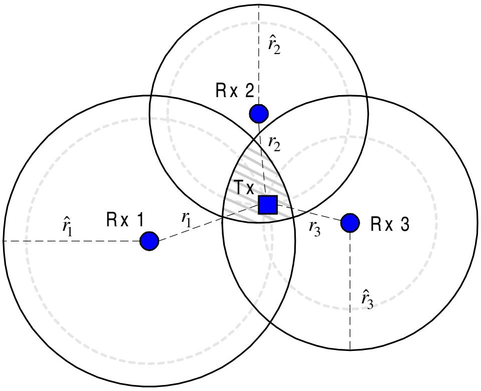


Fig. 3. Time of Arrival geolocation method.

Once TOA is measured, the distance between the transmitter and receiver can be simply determined since the propagation speed of the radio signal is approximately the speed of light $c = 3 \times 10^8 \text{ m s}^{-1}$. The estimated distance at the receiver will geometrically define a circle, centered at the receiver, of possible transmitter positions. TOA measurements at three receivers will provide a position fix and given receiver coordinates and distances from the transmitter to receivers, the transmitter coordinates can be easily calculated. Due to multipath propagation, no-line-of-sight (NLOS) signal path and other impairments, the TOA-based distance estimates are always larger than the true distance between the transmitter and the receiver as illustrated in Fig. 2 where $\hat{r}_1$, $\hat{r}_2$ and $\hat{r}_3$ are the estimated distances and r_1, r_2 and r_3 are the true distances. Three TOA measurements determine a region of possible transmitter position as shown in Fig. 2. A nonlinear least square (NL-LS) method is usually used to obtain the best estimation iteratively by minimizing the estimation errors [6][4]:

$$e_i(x, y) = \hat{r}_i - \sqrt{(X_i - x)^2 + (Y_i - y)} \tag{1}$$

for $i = 1, 2, ..., N$ where (X_i, Y_i) are receiver coordinates and (x, y) is transmitter coordinates. Sometimes the transmission time t_0 of the signals at the transmitter is also taken into account as the third variable:

$$e_i(x, y, t_0) = c(t_i - t_0) - \sqrt{(X_i - x)^2 + (Y_i - y)} \tag{2}$$

where t_i is the receiving time of the signal at the *i*-th receiver. A constrained NL-LS algorithm is also available which makes use of the fact that TOA-based distance estimates are always larger than the true distance [8]. The same as in AOA method, more than three TOA measurements are needed to improve the accuracy of position estimation.

Instead of using TOA measurements, time difference measurements can also be employed to locate the receiver position. A constant time difference of arrival (TDOA) for two receivers defines a hyperbola, with foci at the receivers, on which the transmitter must be located. Three or more TDOA measurements provide a position fix at the intersection of hyperbolas. NL-LS method can also be used to obtain the best estimation of the transmitter position by minimizing the estimation error:

$$e_{i,j}(x,y) = c\hat{\tau}_{i,j} - \left[\sqrt{(X_i - x)^2 + (Y_i - y)^2} - \sqrt{(X_j - x)^2 + (Y_j - y)^2}\right] \qquad (3)$$

for $i, j = 1, 2, ..., N$ where $\hat{\tau}_{i,j}$ is the TDOA measurement of *i*th and *j*th receivers. There are some other methods to solve the hyperbolic position estimation problem as proposed in [12], [13], [14] and [15]. Compared to TOA method, the main advantage of TDOA method is that it does not require the knowledge of the transmit time from the transmitter while TOA method requires. As a result, strict time synchronization between transmitter and receivers is not required. However, TDOA method requires time synchronization among all the receivers.

2.4 Received Signal Strength

If the power transmitted by mobile terminal is known, measuring received signal strength (RSS) at receiver will provide the distance between the transmitter and the receiver using a known mathematical model for radio signal path loss with distances. The same as in the TOA method, the measured distance will determine a circle, centered at the receiver, on which the mobile transmitter must lie. Three RSS measurements will provide a position fix for the mobile. Due to shadow fading effects, RSS method results in large range estimation errors. The accuracy of this method can be improved by utilizing pre-measured received signal strength contour centered at the receiver [16]. A fuzzy logic algorithm was shown in [17] to be able to significantly improve the location accuracy.

2.5 Received Signal Phase

Signal phase is another possible geolocation metric. It is well known that with the aid of reference receivers to measure the carrier phase, differential GPS (DGPS) can improve the location accuracy from about 20m to within 1m compared to the standard GPS, which only uses pseudorange measurements [7]. One problem associated with the phase measurements lies in the ambiguity resulted from the periodicity (with period 2π) of the signal phase while the standard pseudorange measurements are unambiguous. Consequently, in the DGPS, the ambiguous carrier phase measurement is used in fine-tuning the pseudorange measurement. A complementary Kalman filter is used to combine the low noise ambiguous carrier phase measurements and the unambiguous but noisier pseudorange measurements [7]. For indoor geolocation system, it is possible to use the Received Signal Phase method together with TOA/TDOA or RSS method to fine-tune the location estimate. However unlike the application scenario of DGPS where LOS signal path is always observed, the multipath and no-line-of-sight condition of the indoor radio channel causes more errors in the phase measurements.

3. Example Wireless Indoor Geolocation Systems

Commercial indoor geolocation products have already appeared in the market. In this section, we present a technical overview of two example systems, Pinpoint Local Positioning System and Paltrack indoor geolocation system. Both companies claim that the indoor geolocation systems they developed can provide adequate location accuracy and location services in indoor environment.

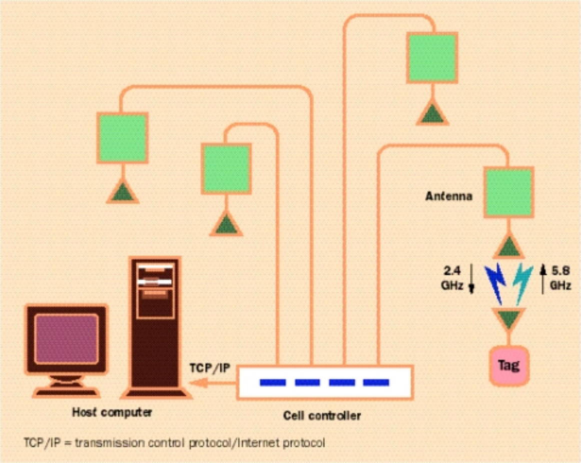


Fig. 4. PinPoint system architecture [4].

3.1 PinPoint Local Positioning System [4]

The system architecture of the PinPoint local position system is shown in Fig. 4. The PinPoint system uses simple-structured tags that can be attached to valuable assets or personnel badges. Indoor areas are divided into cells while each cell being served by a Cell Controller. The Cell Controller is connected to at most 16 antennas located at known positions. To locate tag position, Cell Controller transmits 2.4 GHz spread spectrum signal through different antennas in TDD mode. Once receiving signals from the Cell Controller antenna network, tags simply change the frequency of the received signal to 5.8 GHz and transmit back to the Cell Controller with tag ID information phase-modulated onto the signal. The distance between tag and antenna is determined by measuring round trip time of flight. With the measured distances from tag to antennas, the tag position can be obtained in the same way as in TOA method. A Host Computer is connected to Cell Controller through TCP/IP network to manage the location information of the tags. Since the Cell Controller generates the signal and measures round trip time of flight, there is no need to synchronize the clocks of tags and antennas.

The multipath effect is one of the limiting factors for indoor geolocation. Without multipath signal components, the time of arrival (TOA) could be easily determined from the triangular auto-correlation function of the spread spectrum signal. The triangular auto-correlation peak is two chips' (clock periods) wide at its base, and the time to rise from the noise floor to the peak is one chip. If the chipping rate were 1 MHz, it would take 1000 ns to rise from the noise floor to the peak, providing a "ruler" with a thousand 30-cm increments. A 40-MHz chipping rate was chosen for PinPoint system, providing a ruler of 25 ns that provides real-world increments of about 3.8 meters. Because of regulatory restrictions in the 2.44 and 5.78 GHz bands, faster chipping rates are not easy to achieve, and signal-processing techniques must be used to further improve the accuracy. Also to minimize the multipath effect, different frequency bands are used for uplink and downlink communication to avoid interference between the channels.

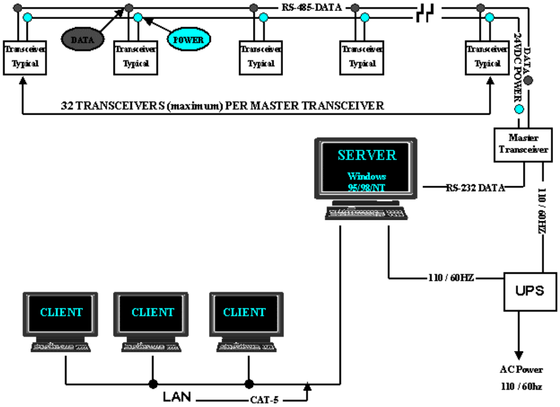


Fig. 5. Paltrack system architecture [5].

3.2 PalTrack Indoor Geolocation System [5]

The infrastructure of Paltrack indoor geolocation system, developed by Sovereign Technologies Corp., consists of tags, antennas, cell controllers and administrative software Server system as shown in Fig. 5. The PalTrack system utilizes a network structure that resides on an RS-485 node platform. A network of transceivers is located at known positions within the serving area while the transmitter tags are attached to assets. The tag transmitters transmit unique identification code at 418 MHz frequency band to a network of transceivers when on motion or at predefined time intervals. Transceivers estimates the tag location by measuring received signal strength (RSS) and utilizing a robust RSS-based algorithm patented by Sovereign Technologies Corp. The Master Transceiver collects measured information from the transceivers and relays it to a PC-based Server system. The accuracy for PalTrack is

0.6 to 2.4 m. The key component of the PalTrack system is the RSS-based geolocation algorithm.

4. An Indoor Geolocation Demonstrator

As we mentioned earlier, the indoor radio channel is very different from that of GPS systems or cellular system. To study the performance of various indoor geolocation techniques, a software indoor geolocation demonstrator is being developed at CWINS, WPI. Some general descriptions of the demonstrator as well as some simulation results will be presented in this section.

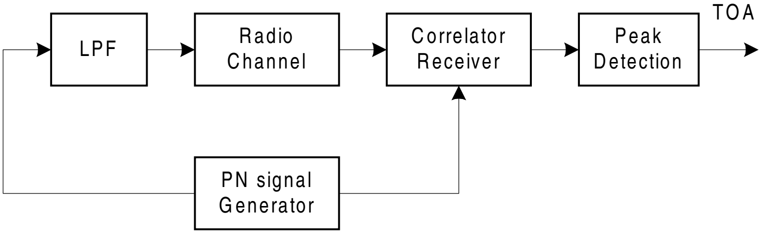


Fig. 6. Simulation of baseband DSSS geolocation system.

4.1 DSSS Indoor Geolocation System

The direct-sequence spread spectrum (DSSS) technique has been used in ranging systems for many years and is the principle behind GPS techniques. A block diagram of the simulated baseband DSSS geolocation system is shown in Fig. 6. For convenience, the PN (pseudo-random noise) signal generator is used in both transmitter and receiver with the assumption of time synchronization between geolocation transmitter and receiver. The lowpass filter (LPF) is used to take into account the band-limitation condition of the realistic radio transceiver systems. The autocorrelation characteristics of the PN sequence are fundamental to the distance (or TOA) estimation. A 31-chip Gold sequence was used with the chipping interval $T_c = 25\,\text{ns}$ and the sampling period $T_s = 5\,\text{ns}$ which provides a width of 50ns at the base of the triangular autocorrelation function of PN signal and an accuracy of 5ns in the TOA measurements. While the multipath radio channel spreads the transmitted signal, geolocation receivers are only interested in detecting the DLOS path, i.e. determining the arrival time of the first peak in the output autocorrelation signal [1]. With TOA measurements of signals from multiple reference transmitters, the position estimate of the receiver can be obtained iteratively using nonlinear least square algorithm.

4.2 Channel Measurement and Ray-Tracing

The channel profiles used in this simulation can be obtained in two ways. One is using a frequency domain measurement system described in [18]. The centerpiece in this system is a network analyzer that sweeps the channel from 900- 1100 MHz. The output signal is first amplified with an amplifier and then connected to the transmitter antenna through a long cable. The receiver antenna passes the signal through a chain of low noise amplifiers that are connected to input port of the network analyzer. The network analyzer records the frequency magnitude and phase responses of the channel. Fig. 7 shows an example of measured radio channel frequency response. For the geolocation simulation, we need to measure the actual frequency-domain channel response between the points of each reference transmitter and receiver. Then the channel impulse responses between transmitter and receiver, which we refer to as a channel time profile or simply a profile, can be obtained by taking the Fourier transform to the measured frequency channel response.

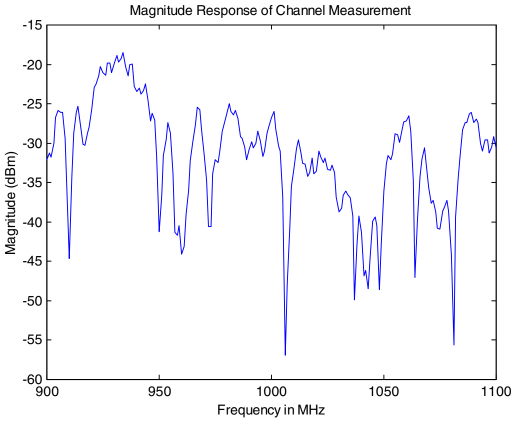


Fig. 7. Magnitude response of a frequency domain channel measurement.

Another way to obtain the channel profiles is using ray-tracing channel modeling method. Radio signals with frequencies larger than 300 MHz have extremely small wavelengths compared to the dimensions of building features so that electromagnetic waves can be treated simply as rays [19]. This is the principle behind ray-tracing method for radio channel modeling. In our simulation, we used CWINS 2D ray-tracing software to obtain the channel profiles, as shown in Fig. 8, between each reference transmitter and the receiver. Then the channel profiles can be directly used in geolocation simulations to obtain TOA and distance measurements.

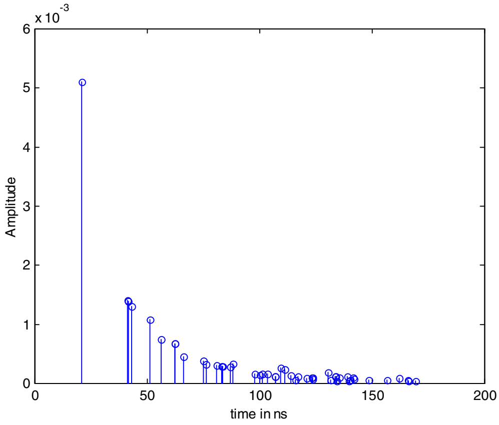


Fig. 8. Channel profiles obtained using 2-D ray-tracing software.

4.3 Simulation Results

A fundamental issue in geolocation is the analysis of the accuracy of positioning. For spread spectrum geolocation systems, one of the limiting factors is the available channel bandwidth, i.e. the larger the channel bandwidth, the higher the measurement resolution which closely relates to the accuracy of TOA measurement. The multipath indoor radio channel makes the analysis very complicated. This section presents results of simulations that relate the bandwidth with the accuracy of positioning.

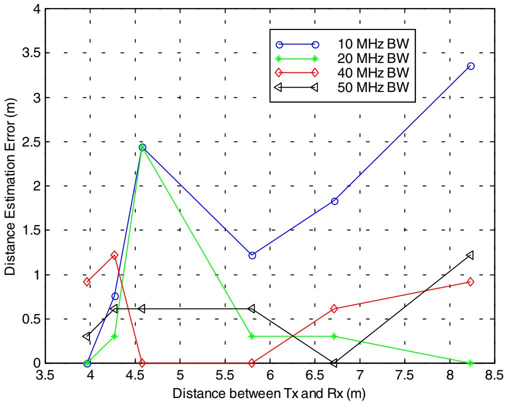


Fig. 9. Effects of channel bandwidth in ranging errors (with measurement-based method) [20].

Fig. 9 shows results obtained from channel measurement-based simulations. Table 1 shows the ranging errors between the transmitters and the receiver for different signal bandwidth obtained from both measurement based and ray-tracing based simulations. In Fig. 9 we observe that ranging errors are less than 3.5 m in all the cases. But to achieve less than 1.5 m accuracy, a bandwidth of larger than 20 MHz is needed.

Table 1. Mean ranging errors using measurement based and ray-tracing based methods [20][21].

Bandwidth (MHz)	Measurement based ranging error (m)	Ray-tracing based ranging error (m)
10	2.18	1.48
20	1.10	0.89
40	1.22	0.55
50	1.07	0.32

By using nonlinear least square algorithm to estimate receiver position from TOA measurements, we can employ more than two TOA measurements. Table 2 shows mean of the location estimation errors when different numbers of TOA measurements are used in the geolocation algorithm. Both measurement-based and ray-tracing based simulations show consistent results. But the ray-tracing based method is more convenient since for the measurement-based simulations, channel responses between all the transmitters and the receiver have to be measured.

Table 2. Mean location estimation errors with different numbers of TOA measurements [21].

Number of TOAs used in geolocation	Measurement based location error (m)	Ray-tracing based location error (m)
3	1.25	0.81
4	0.69	0.52
5	1.05	0.31
6	1.05	0.49
7	1.21	0.39

5. Conclusions

In this paper, we briefly reviewed various geolocation metrics and discussed usability of these metrics in indoor environment. Due to the serious multipath and no-line-of-sight propagation condition of the indoor radio channel, TOA/TDOA and Received Signal Strength methods are more appropriate than AOA method. Technical overview of two example wireless indoor geolocation products is then presented. Simulation results obtained from a software indoor geolocation testbed show that available channel bandwidth plays an important role in the accuracy of indoor geolocation systems. It was shown that ray-tracing based method provides consistent results with that of measurement-based method and the ray-tracing based method is more convenient than the frequency-domain measurement method.

Acknowledgement

The authors would like to express their appreciation to TEKES, Nokia, and Finnish Airforce for supporting most parts of this project. We are thankful to two MQP groups at CWINS, WPI for their contributions to this work: K.H. Shah, C.M. Kelly and D.S. Hastings for measurement-based simulations and S. Dasmah, C.D. Le and T.Q. Nguyen for ray-tracing based simulations. We also thank Dr. Jacques Beneat and Dr. Prashant Krishnamurthy, our colleagues at CWINS, for fruitful discussions and a variety of help.

References

[1] K. Pahlavan, P. Krishnamurthy and J. Beneat, “Wideband radio propagation modeling for indoor geolocation applications”, IEEE Comm. Magazine, pp. 60-65, April 1998.
[2] P. Krishnamurthy, K. Pahlavan, J. Beneat, “Radio propagation modeling for indoor geolocation applications”, Proceedings of IEEE PIMRC'98, September 1998.
[3] Jay Werb and Colin Lanzl, “Designing a positioning system for finding things and people indoors”, IEEE Spectrum, vol. 35, No. 9, Sep. 1998.
[4] PinPoint Local Positioning System, http://www.pinpointco.com/.
[5] PalTrack Tracking Systems, http://www.sovtechcorp.com/.
[6] J. Caffery, Jr. and G.L. Stuber, “Subscriber Location in CDMA Cellular Networks”, IEEE Trans. Veh. Technol., vol. 47, No. 2, May 1998.
[7] E.D. Kaplan, Understanding GPS: Principles and Applications, Artech House Publishers, 1996.
[8] G. Morley and W. Grover, “Improved location estimation with pulse-ranging in presence of shadowing and multipath excess-delay effects”, Electronic Letter, vol. 31, pp 1609-1610, Aug., 1995.
[9] G. Turin, W. Jewell and T. Johnston, “Simulation of urban vehicle-monitoring systems”, IEEE Trans. Veh. Technol., vol. VT-21, pp. 9-16, Feb. 1972.
[10] H. Hashemi, “Pulse ranging radiolocation technique and its application to channel assignment in digital cellular radio”, Proc. IEEE VTC’91, pp. 675-680, 1991.
[11] P. Goud, A. Sesay and M. Fattouche, “A spread spectrum radiolocation technique and its application to cellular radio”, Proc. IEEE Pacific Rim Conf. Comm., Comp. and Signal processing, 1991, pp. 661-664.
[12] W.H. Foy, “Position-location solutions by Taylor-series estimation”, IEEE Trans. Aerospace and Electronic Systems, vol. AES-12, pp. 187-194, Mar. 1976.
[13] D.J. Torrieri, “Statistical theory of passive location system”, IEEE Trans. Aerospace and Electric Systems, vol. AES-20, No. 2, Mar. 1992.
[14] J.S. Abel and J.O.Smith, “A divide-and-conquer approach to least-squares estimation”, IEEE. Trans. Aerospace and Electric Systems, vol. 26, pp. 423-427, Mar. 1990.
[15] Y.T. Chan and K.C. Ho, “A simple and efficient estimator for hyperbolic location”, IEEE Trans. Signal Processing, vol. 42, No. 8, pp. 1905-1915, Aug. 1994.
[16] W. Figel, N. Shepherd and W. Trammell, “Vehicle location by a signal attenuation method”, IEEE Trans. Vehicular Technology, vol. VT-18, pp. 105-110, Nov. 1969.
[17] Han-Lee Song, “Automatic Vehicle Location in Cellular Communications Systems”, IEEE Trans. Vehicular Technology, vol. 43, No. 4, pp. 902-908, Nov. 1994.
[18] S.J. Howard and K. Pahlavan, “Measurement and analysis of the indoor radio channel in the frequency domain”, IEEE Trans. Instr. Meas. No. 39, pp. 751-755, 1990.
[19] K. Pahlavan and A. Levesque, Wireless Information Networks, John Wiley and Sons, New York, 1995.

[20] K.H. Shah, C.M. Kelly and D.S. Hastings, MQP Project Report: Wireless Indoor Geolocation System, CWINS, Worcester Polytechnic Institute, May 1999.

[21] S. Dasmah, C.D. Le and T.Q. Nguyen, MQP Project Report: Simulation Platform for Performance Evaluation of Indoor Geolocation, CWINS, Worcester Polytechnic Institute, January 2000.

Priority Based Multiple Access for Service Differentiation in Wireless Ad-Hoc Networks

u ng[1] n r him n ou[2]

[1] Department of Electrical Engineering, National University of Singapore,
10 Kent Ridge Crescent, Singapore 119260
engp8843@nus.edu.sg

[2] Centre for Wireless Communications, National University of Singapore,
20 Science Park Road, #02-34/37, Singapore 117674
brahim@cwc.nus.edu.sg

Abstract. This article describes the Priority Based Multiple Access (PriMA) protocol, a new medium access control (MAC) protocol for single-channel ad-hoc networks. Unlike previously proposed protocols, PriMA takes into account the QoS requirements of the packets queued in stations to provide each station with a priority-based access to the channel. The direct support of PriMA for ad-hoc routing is that when some stations act as hubs in the routing structure and route packets for other stations besides their own, they can have high priorities and obtain larger share of bandwidth. Simulation results show the potential benefits that PriMA brings about to ad-hoc networks, and confirm PriMA as an initial step towards QoS provision in ad-hoc networks.

1 Introduction

n ho n twork i n mi multi hop wir l n twork th t i t li h group o mo il t tion without th i o n pr xi ting n twork in r tru tur or ntr liz mini tr tion. t n in t ll qui kl in m r g n or om oth r p i l itu tion n i l on gur l whi h m k it v r ttr tiv in m n ppli tion . om ppli tion in lu r mot n ing (.g. rthqu k r nvironm nt l t g th ring) ro oti ommuni tion (.g. l ning r ghting p trolling o ni hing rming) nt rn t n L O on t ll tion . nt rn t ngin ring k or ( ) h t up workgroup n m Mo il ho twork (M) 1 to v lop n volv M routing p i tion n intro u th m to th nt rn t t n r tr k. nt i u ion in M m iling li t on M ppli tion n rio n giv n i o th l wh n M n plo (.g. n ho n twork o w ork t xi).

n i nt m ium ontrol (M) proto ol through whi h mo il t tion n h r ommon ro t h nn l i nti l in n ho n t work u th m ium or h nn l i r r our . u to th limit tr n mi ion r ng o mo il t tion multipl tr n mitt r within r ng o th

C.G. Omidyar (Ed.): MWCN 2000, LNCS 1818, pp. 14–30, 2000.

m r iv r m not know on noth r tr n mi ion n thu in ff t “hi n” rom on noth r. h n th tr n mitt r tr n mit to th m r iv r t roun th m tim th o not r liz th t th ir tr n mi ion olli t th r iv r. hi i th o ll “hi n t rmin l” pro l m 2 whi h i known to gr throughput igni ntl . u to th ir multi hop h r t ri ti ho n twork uff r mu h mor rom th hi n t rmin l pro l m th n wir l L o.

o r th hi n t rmin l pro l m v riou i tri ut M proto ol w r propo in th lit r tur . M (Multipl olli ion voi n) proto ol 3 propo x h ng o hort qu t to n ( ) n l r to n ( ) p k t tw n p ir o n r n r iv r or tu l t p k t tr n mi ion n orm th i or v r l oth r mor ophi ti t h m . mong th m M (loor qui ition Multipl with on p r i t nt rri r n ing) 4 i immun to th hi n t rmin l pro l m n n hi v goo throughput in ho n twork . 02.11 ommitt l o propo M proto ol ll i tri ut oun tion ir l M ium ontrol (M) or wir l ho L whi h in n i v ri nt o M / proto ol . M provi i n / m tho . hh n upt 6 h v hown th t th p r orm n o / m tho gr mu h low r th n th i m tho wh n th num r o hi n t rmin l i l rg or th off r lo i igni ntl l rg r th n th h nn l p it th r or th / m tho i mor ro u t to flu tu tion in p r m t r v lu whi h r ommon in ho n twork . n M / m tho 4 w i log in lu ing i u to om t th hi n t rmin l pro l m. M how v r till nnot pr v nt t p k t rom olli ing with ontrol p k t (n) n oth r t p k t in t it u ophi ti t mo i in r xpon nti l koff h m to qui kl r olv olli ion n thu in r throughput.

h M proto ol p r orm w ll in olving hi n t rmin l pro l m how v r non o th m t k n t p tow r provi ing p k t l v l o p r m t r u h p k t lo r tio p k t l t . h n upport onl t ffort liv r rvi thu limit th ppli tion o ho n twork . ntl th r h n urg in mo i ing n xt n ing qu lit m ium h m to upport priorit wh n th p k t qu u t t tion h v i r nt o r quir m nt . lthough thi r r h ffort o not omp t with th ffort plo in th ir l L (in r tru tur) om r r h r h v n working on o upport or i tri ut wir l n twork . n t tion with r l tim p k t in th qu u woul j m th h nn l with l k ur t ( ) who l ngth i proportion l to th l in urr . h t tion th t n th long t win to th h nn l n n tr n mit it p k t th r t r. ow v r thi ppro h il wh n hi n t rmin l xi t tho hi n t rmin l m h v xp ri n th m l n h on t ntion p rio i not gu r nt to pro u uniqu winn r thu r l tim t p k t will till uff r rom olli ion . n M (roup llo tion Multipl) proto ol w propo or h uling r l tim n t gr m tr in ingl hop wir l ho L . M in lu ont ntion p rio uring whi h t tion n n r qu t to join tr n mi ion group n ont ntion r p rio uring whi h t tion in tr n mi ion group t k turn to tr n mit p k t . hi ppro h o not work w ll i hi n t rmin l xi t . hi n t r min l o not join th tr n mi ion group th t it m int r r with M nnot n ur t p k t to r rom olli ion. hi n t rmin l o join th tr n mi ion group to voi olli ion ollowing th m logi th n ll th oth r t tion in th n twork h v to join th m tr n mi ion group on on . t i v r i ult to m int in th glo l group u to th n mi n tur o ho n twork. n ition w nnot n t rom p ti l r u . h r or th proto ol th t w r origin ll propo or wir l L r not ir tl ppli l to multi hop ho n twork . li v th t uit l M proto ol or ho n twork houl r oth hi n t rmin l pro l m n o i u t th m tim .

noth r motiv tion or priorit i to provi tt r upport or ho routing. On on h n tho proto ol origin ll propo or wir l L norm ll o not t k routing into ount th xp t wir l point th t n r h ll oth r t tion n r l p k t or th m to plo . h r or th r mor t or n twork with in r tru tur th n ho n twork wh r routing i noth r m jor i u . On th oth r h n unlik in onv ntion l wir n twork n ho t th t rout r in ho n twork norm ll h onl on n twork int r th r r no p r t link or th m to rout p k t or x h ng routing in orm tion. p i ll wh n om t tion in n ho n twork h v lu t r h 9 or long to th or o th rout ing tru tur 10 mor tr will tr n it through th m. O viou l th n high r prioriti in ing th h nn l to rout p k t or oth r in ition to th ir own.

n thi p p r w propo n w proto ol n m riorit Multipl (riM) to upport iff r nti t prioriti to th h nn l. t impl m nt M l v l knowl gm nt in M whil opting th olli ion r t tr n mi ion h r t ri ti o M . Mor import ntl it impl m nt nov l i tri ut h uling lgorithm whi h giv t tion n mi priorit to th h nn l t king into ount oth th p k t l r quir m nt n th p k t lo r tio in urr th ongoing ion. h r m in r o th p p r i org niz ollow . tion 2 ri riM proto ol in t il. tion 3 omp r imul tion th p r orm n o M 02.11 M n riM . tion 4 on lu thi p p r.

2 PriMA Protocol

2.1 Overview

riM proto ol r quir t tion th t wi h to n t p k t to quir th h nn l or tr n mitting th p k t. h h nn l i quir t li h ing n i log tw n th n r n th r iv r. lthough multipl ontrol p k t m olli t p k t r lw nt r rom olli ion . hi i hi v th n or r quir m nt o iz r l tion hip tw n n w ll iff r nt p rio t tion houl w it t r r iving p k t or n ing th h nn l u . ullm r t l. 11 h v giv n t il ription out thi . h how th t long th l ngth o p k t i u i ntl long r th n th t o th p k t n t j mming ig n l to pr v nt oth r t tion rom tr n mi ion. t tion th t h r th h nn l u houl w it long nough or th po i l ongoing t tr n mi ion to go on uno tru t . riM trivi l xt n ion to thi i to r pl n knowl g m nt p k t t r u ul r ption o t p k t. hi i l o th ommon in om oth r M proto ol lik M . h xt n ion o M l v l knowl gm nt till n ur t tr n mi ion r rom olli ion .

h mo t import nt tur o riM i th t th to th h nn l i on prioriti o th p k t qu u in t tion . hi i hi v thr tim r l ul t or ing to th o r quir m nt o th t p k t in h t tion. h r t i th *access timer* th t t tion houl w it t r th h n n l om i l or tr n mitting n p k t. h tim i imil r to (i tri ut nt r r m p) in M n th il nt p rio o h nn l in M . ow v r unlik th x l ngth o n th r n om l ngth o th il nt p rio in M th tim i n mi ll ju t on th o r quir m nt . h on i th *delay timer* rri in v r p k t. t in i t how long th int n r iv r n w it or r pl ing with p k t. h r or n rli r nt p k t m pr mpt l t r nt p k t whi h in i t high r priorit t p k t tr n mi ion r qu t. h thir i th *backoff time* th t t tion houl w it or r tr n mi ion wh n olli ion o ur . h koff tim i uni orml i tri ut how v r th upp r oun o th i tri ution v r mong t tion . t tion th t hol high r priorit p k t h v low r upp r oun . h r or th n t ti ti ll r ov r rli r th n oth r t tion n i or th h nn l g in. or t tion qu u with norm l t p k t w ju t t l rg upp r oun omp r l to th t o M . h r or our koff h m i not n ril mor pron to oll p th n in r xpon nti l koff with upp r oun .

2.2 Description of PriMA

o impli our ription o riM th pro ing tim n tr n mit to r iv turn roun tim r ignor . t tion “ t t ” olli ion i it n th h nn l u without ing l to r iv n int lligi l p k t. ollowing w n om o th not tion u in thi tion

rtPacket t p k t with o p r m t r p i
nrtPacket t p k t without n o p r m t r p i
T_d M ximum on hop h nn l prop g tion l

T_{type} im to tr n mit p k t o t p *type* wh r *type* n ith r or

T_{delay} llow l tim to r pl to rri in p k t

T_{left} im l t or *rtPacket* to liv r wh n T_{left} i l th n thr h ol th orr pon ing *rtPacket* will ropp .

T_{max} M ximum tim to ompl t on u ul tr n mi ion

T_{access} im r quir to n th h nn l i l or tr n mitting n p k t

T_{defer} h long t tim t tion houl r wh n “ t ting” olli ion or nt ring koff. t qu l $T_{data} + 3\ \ T_d$.

T_{unit} im u tor to m p p k t l r quir m nt to koff tim r

N_{lost} um r o p k t ropp uring th ion

N_{sent} um r o p k t nt uring ion

PLR k t Lo tio (o r quir m nt).

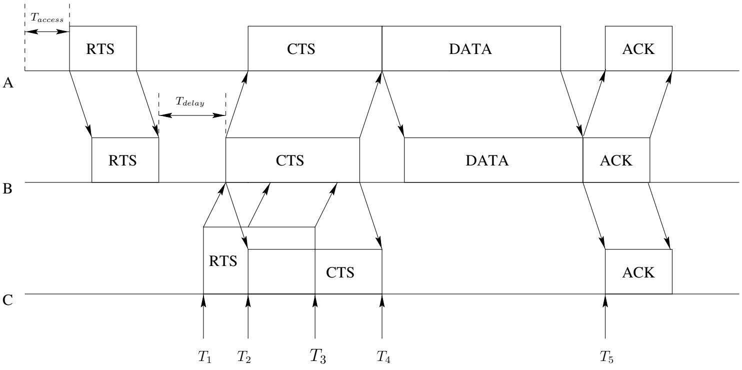


Fig. 1. riM illu tr tion

igur 1 how how riM op r t . n thi gur t tion n r hi n rom h oth r n t tion i n igh or o oth n . in ll oth r olli ion voi n multipl t hniqu riM u th om in tion to impl m nt olli ion voi n i to n ur th pro l m o hi n t rmin l i r qu t l riM impo M on ition on th iz o th n p k t 11

- $T_{rts} > 2\ \ T_d$
- $T_{cts} > T_{rts} + 2\ \ T_d$

n t tion th t n th h nn l u houl w it or T_{defer} to l t th po i l ongoing t tr n mi ion to ni h uno tru t . M r tion l in impo ing th on ition i th t lthough i hi n rom n thu i not h r i n n p k t th t t mo t olli with th r pl rom to th on ition n ur th t h r th tr il r o th n thu woul ort n tr n mi ion n nt r th koff pro ur . hi n ur th t th t p k t rom to i tr n mitt olli ion r . ot th t it i not n r to opt th r tri tion (n or in M) in riM to provi o t o . v rth l w li v th t it o not m k n to r th pro l m o o (it o t o) without r ing r t th hi n t rmin l pro l m p i ll wh n th off r tr lo n v r high. riM ppro h to provi ing o n l o ppli to 02.11 M .

n riM n t tion th t h p k t to n houl w it until it n th h nn l i l or rt in mount o tim ll *access timer* or tr n mitting n r qu t. h tim l ul tion hown in lgorithm 1 i on th r quir o n th p r iv o .

Algorithm 1 tim r l ul tion

```
1:  if (nrtPacket or rtPacket with $T_{left} > 1$ sec) then
2:    $T_{access} = 2 -T_d$
3:  else
4:    if ($T_{left} > 0$) then
5:      $T_{access} = T_{left} - (N_{lost} - PLR -N_{sent}) -T_{max}$
6:      if ($T_{access} < 0$) then
7:        $T_{access} = 0$
8:      else
9:        $T_{access} = 2 -T_d/-\log_2 (T_{access}/2)-$
10:     end if
11:   end if
12: end if
```

ot th t in or r to r u th w t o n wi th u to thi tim tim n itiv p k t whi h h v tim to liv o mor th n on on om p t t th m priorit l v l th non tim n itiv p k t . h p k t woul g in high r prioriti wh n th om mor urg nt. ot th t initi ll t th p k t r t tt mpt T_{left} i initi liz to v lu proportion l to th m ximum tol r l l o th p k t. l o throughout th lgorithm T_{left} r with th ti k o th lo k in th m proportion. h l ul tion in lin o th lgorithm how th t th mor h t tion uff r x roppp p k t ($N_{lost} - PLR -N_{sent}$) th hort r T_{access} will . hi x rpt o p u o o l rl how th t wh n two t tion tt mpt to th h nn l imult n ou l T_{access} will iff r nti t th ir tt mpt or ing to th ir p k t o r quir m nt.

h n t tion u in it to th h nn l it t “ l ” l in it p k t n n it out. h v lu o th l i not T_{delay} whi h in i t how long th int n r iv r n w it or r pl ing with p k t. T_{delay} i l ul t hown in lgorithm 2.

Algorithm 2 l tim r l ul tion

```
if (nrtPacket) then
    $T_{delay} = 2 -T_{rts}$
else
    if ($T_{left}/T_{max} > 2$) then
        $T_{delay} = 2 -T_{rts}$
    else
        $T_{delay} = (T_{left}/T_{max}) -T_{rts}$
    end if
end if
```

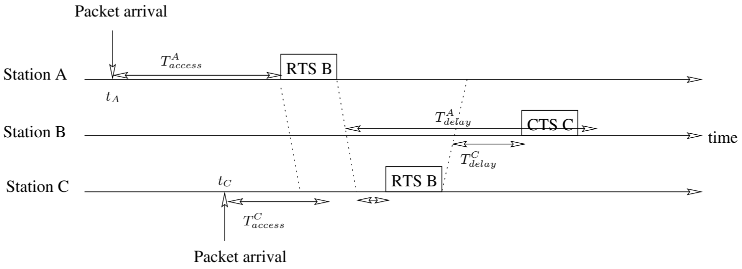


Fig. 2. llu tr tion o th pr mption in riM

o illu tr t th import n o thi on tim r l t u on i r th x mpl in igur 2 wh r two t tion A n C tt mpt to ommuni t with t tion B. will n gl t th prop g tion tim in thi x mpl . t p k t rriv t th M l r t tim t_C l t r th n t p k t who rriv t t_A how v r l t rg t i mor tring nt th n . u to th l k o glo l oor in tor in th t m n th l k o glo l t t th r i no w or to u in n ing n p k t to or o in $t_C + T^{C}_{accces} > t_A + T^{A}_{access}$. t v lu T^{A}_{delay} in it p k t whi h t ll th mount o tim it n w it or r pl ing with . t r h r th n o (w will l t r th wh r n r hi n rom h oth r) it will ppl th m pro ur it u or th r t how v r with onl th r m ining tim r. n oth r wor will w it until it n n i l h nn l or T_{left} $t_C + T^{C}_{accces} - t_A - T^{A}_{access}$ t rting t $t_A + T^{A}_{access} + T_{RTS}$. thi l t tim i m ll r th n T^{A}_{delay} i. . i tim r xpir or r pli with to th n n n to n th t will pr mpt . hi w lthough u in n ing th or thi l tt r h th po i ilit to pr mpt or it t rt tr n mitting it t p k t.

n th l ul tion o T_{delay} or r l tim p k t onl T_{left} i u . hi i u T_{left} i ompo it v lu ff t oth p k t l n p k t rop r tio. hu w n k p th lgorithm impl t ff tiv . hi lgorithm how th t th p k t or t p k t th t i mor l n itiv will h v hort r T_{delay} wh n th t p k t i out to xpir . h l ul tion will p rmit th p k t to g in high r prioriti t r th r l or om tim . t tion u u ll o not r pl with imm i t l t r r iving n p k t unl r quir to o o th r or t tion m r iv multipl p k t in row. th l t r rriv p k t r quir hort r l th n th t o th rli r on th t tion n r pl to it r t. hi m k it po i l or l n itiv t p k t to pr mpt oth r t p k t . n ition T_{delay} o not x two tim th p k t tr n mi ion tim to minimiz th proto ol ov rh .

t i in vit l th t ontrol p k t m olli with h oth r th r or tho t tion th t t t olli ion houl l o w it or T_{defer} n th n k off or r n om tim $T_{backoff}$ whi h i l ul t hown in lgorithm 3 wh r $U(0,x)$ i uni orml i tri ut r n om num r in th int rv l $(0,x)$.

Algorithm 3 koff tim r l ul tion

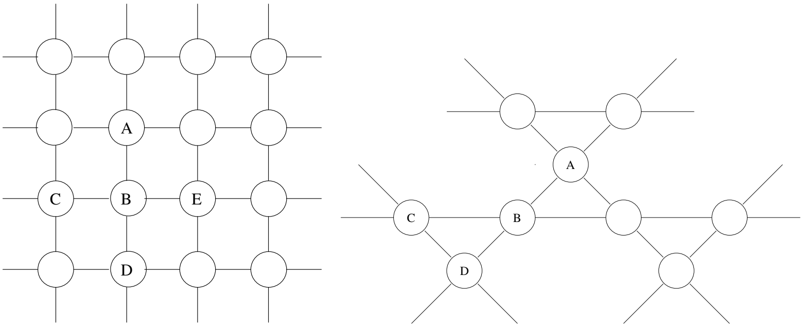
```
if (nrtPacket or rtPacket with $T_{left}/T_{unit} > Maxtimer$) then
    $T_{backoff} = U(0, Maxtimer) - 2T_d$
else
    $T_{backoff} = U(0, T_{left}/T_{unit}) - 2T_d$
end if
```

h v lu o *Maxtimer* i impl t to l rg v lu 00 whi h i omp r l to th m ximum tim r u in M . or x mpl th 02.11 ir l L t n r p i tion or ir t qu n pr p trum ( ) u tim r (or *Contention Window* in 02.11 t rminolog) r nging rom 31 to 1023. n riM $T_{backoff}$ i on uni orm i tri ution who upp r oun v lu v ri or ing to th l r quir m nt o t p k t. hi l ul tion will t ti ti ll giv t tion th t h v l n itiv t p k t hort r $T_{backoff}$. hu on v r g th t tion will n th koff p rio rli r th n oth r t tion n i or th h nn l g in. o illu tr t how lgorithm 3 ontri ut in iff r nti ting th t tion l t u r turn to th x mpl o igur 2. n th wh r n r hi n rom h oth r. nnot h r th p k t o n thu t r xh u ting it tim r T^{C}_{access} it woul l o n n p k t whi h woul olli t with p k t. t r T^{A}_{delay} (r p tiv l T^{C}_{delay}) t tion (r p tiv l t tion) noti th olli ion tim out n thu th oth ppl th ov koff lgorithm with th ir r p tiv o t rg t . n thi h high r pro ilit in i ing or th h nn l or o in r w it uni orml i tri ut koff tim rom m ll r int rv l th n o n thu h high r pro ilit to t rmin t it koff or .

rom th ov ription w n th t n mi ll ju ting thr tim r T_{access} T_{delay} n $T_{backoff}$ or ing to th t p k t o r quir m nt riM n giv t tion priorit to th h nn l. t th m tim v riou w iting tim p rio r r ull n to pr v nt ontrol p k t rom olli ing with t p k t .

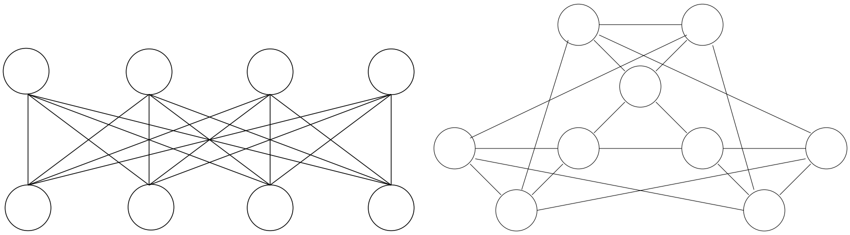
3 Simulation Results

riM i ompl x M proto ol or n l ti l mo lling u o th n mi n tur o th iff r nt tim r thu in thi p p r th p r orm n o riM i inv tig t imul tion . n our xp rim nt w inv tig t m m tri l n twork wh r h t tion h N n igh or n i hi n rom Q n igh or o n on o it n igh or thu h t tion h th m p ti l h r t ri ti . igur 3 how two mpl on gur tion or N 4 Q 2 n Q 3. or x mpl or t tion in gur 3( ) it h 4 n igh or in lu ing t tion n i hi n rom t tion 3 n igh or n . thi gr ph n grow to in nit l l rg w ol or oll p it o th t th tot l num r o t tion in th n twork i limit whil mm tr i m int in . h r ulting gr ph r hown in gur 4.



(a) N=4, Q=3 (b) N=4, Q=2

Fig. 3. wo mpl on gur tion



(a) N=4, Q=3 (b) N=4, Q=2

Fig. 4. wo mpl on gur tion (oll p) um 1M p i l h nn l with z ro pr m l n pro ing ov r h . h v p r orm iff r nt t o imul tion with O Mo l r/ io n w omp r our r ult to tho provi lt rn tiv proto ol no t l M n 02.11 M [1]. l 1 li t th p r m t r u in th imul tion. w ignor th xtr tim in urr h r w r n o tw r th iff r nt nt r r m p ( ) in 02.11 r r u or ingl n th r hown in t l 2.

Protocol	RTS	CTS	DATA	ACK	backoff timer	backoff unit time
FAMA	20-byte	25-byte	500-byte	-	10	160μsec
IEEE 802.11	25-byte	20-byte	500-byte	20-byte	31-1023	6μsec
PriMA	20-byte	25-byte	500-byte	20-byte	800	12μsec

Table 1. roto ol on gur tion p r m t r

DIFS	SIFS	EIFS
12μsec	0μsec	1.3msec

Table 2. xtr on gur tion p r m t r or 02.11

n th r t t o imul tion ll t tion g n r t oi on tr with th m m n r t n ll r quir t ffort liv r rvi . h imul tion m ur th throughput o th proto ol g in t th gr o th no . h r ult r hown in igur . or omp r tiv purpo w l o how th n l ti l r ult or lott LO with p r t knowl g m nt h nn l n on r i t nt M in th gur . h gur mon tr t th t M p r or m n gr r m ti ll wh n th num r o omp ting t tion (in lu ing n igh or n hi n t rmin l) in r . hi i u to th in ff tiv n o uni orm koff h m with m ll tim r in olli ion r olution u in M . ow v r it till p r orm w ll in oth r itu tion u o it immunit to hi n t rmin l pro l m. 02.11 p r orm quit w ll in ll th itu tion u o it mo i in r xpon nti l koff h m . lthough riM p r orm n i not quit out t n ing wh n th num r o omp ting t tion i m ll how v r riM o hi v r th r t l throughput n i omp r l to 02.11 wh n th num r o omp ting t tion in r . hi i not urpri ing. riM intro u ing th iff r nt tim r w ll th M knowl gm nt in t ri throughput to hi v iff r nt prioriti mong th t tion (whi h w how l t r). n ition l rg koff tim r will om mor ff tiv wh n th num r o omp ting t tion in r th r or th throughput iff r n tw n riM n 02.11 i r ing. n oth r

[1] We use its specification for Direct Sequence Spread Spectrum if applicable.

wor riM ptur in t th ff tiv n o M in olving th hi n t rmin l pro l m w ll 02.11 ilit in u t ining throughput t high lo . t ri how v r littl throughput intro u ing th iff r nt tim r in or r to hi v iff r nti tion tw n th t tion hown in th ollowing.

n th on t o imul tion h t tion till g n r t oi on tr with th m m n r t . ow v r w t on t tion (rout r) to g n r t p k t th t r quir l oun liv r whil th oth r t tion (ho t) h v no l r quir m nt. hi in n m k th p k t rom th rout r h v high r priorit th n p k t rom th ho t .

hi n rio m ppli l in wh n om t tion orm group n hoo on o th m to t rout r. n oth r wor th rout r will h n l l rg r mount o tr n thu n om priorit to th h nn l. imul tion r ult r hown in gur 6 n with iff r nt v lu o N n Q. h gur l rl how th t th rout r h high r throughput th n oth r t tion pit th t th t th ll off r th m tr lo to th h n n l. v n wh n th num r o omp ting t tion in r th rout r n till hi v mu h high r throughput th n oth r norm l t tion . hi iff r nti tion nnot hi v M n oth r non prioritiz proto ol u h 02.11 M .

n th thir t o imul tion w l t two t tion (high priorit rout r n low priorit rout r) to g n r t on t nt r t p k t th t r quir iff r nt p k t liv r l oun . inv tig t two p r orm n m tri un r iff r nt n rio n m l th p k t lo r tio (u to xpir tion) n th v r g l . o g t omp r tiv r ult w t th p k t int r rriv l tim qu l to th p k t l r quir m nt oth rwi proto ol lik 02.11 will n v r k p up with th p k t g n r tion r t n v ntu ll will rop n rl ll th p k t . imul tion r ult r hown in gur . h r t thr u gur (( ) ( ) ( )) how th p k t lo r tio g in t th off r lo . h how th t wh n th off r lo in r th p k t rop r tio or 02.11 in r igni ntl whil riM n till m int in low p k t rop r tio or th two rout r . igur ( ) how th t th v r g p k t l hi v th two t p o rout r in riM . h gur l rl how th t wh n th tr lo i v r low th p k t rom ll th t tion xp ri n th m (in igni nt) l . ow v r wh n th tr lo in r th proto ol t rt iff r nti ting th t tion or ing to th ir r quir m nt .

4 Conclusion

riM i n w M proto ol th t i p i ll ign or ho n twork . t n hi v goo throughput in ho n twork wh r hi n t rmin l pro l m i ommon. n ition v r t tion h priorit to th h n n l thu riM n provi l m nt r o upport rom th ottom up m king it goo hoi or upporting high r l r proto ol th t r quir qu lit o rvi . noth r n t o riM i th t it n provi tt r upport or

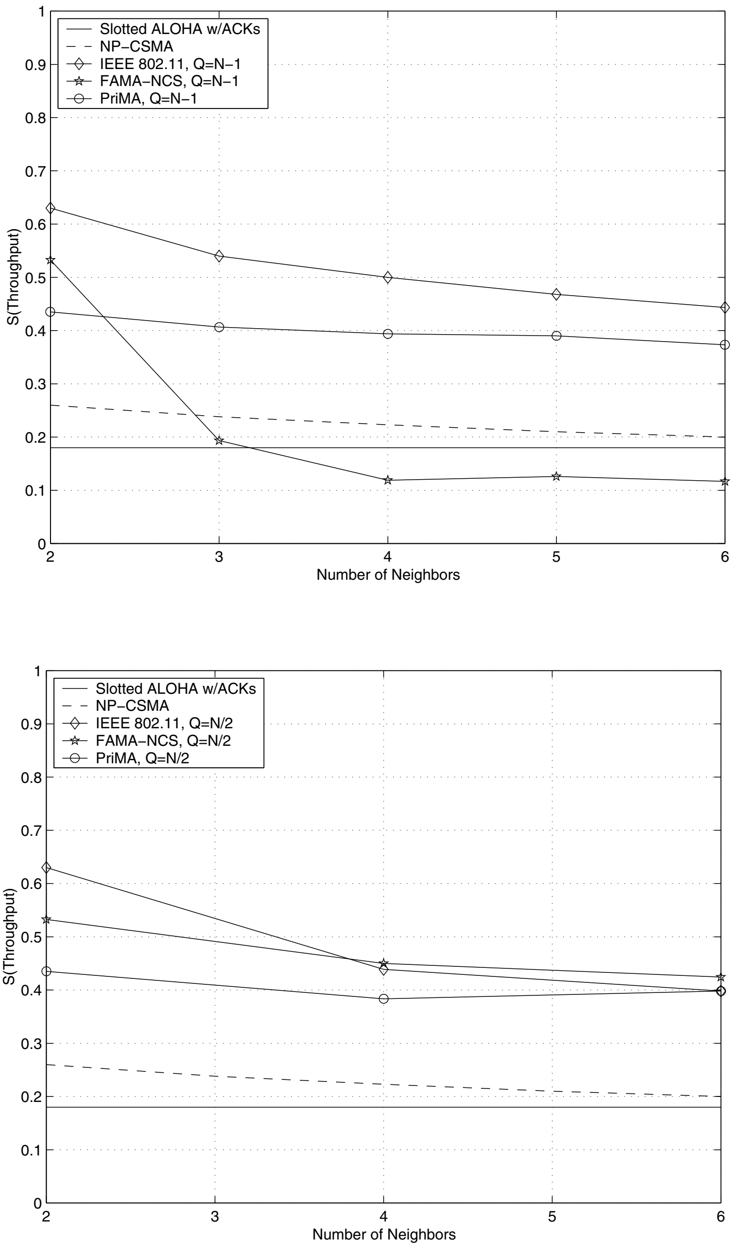
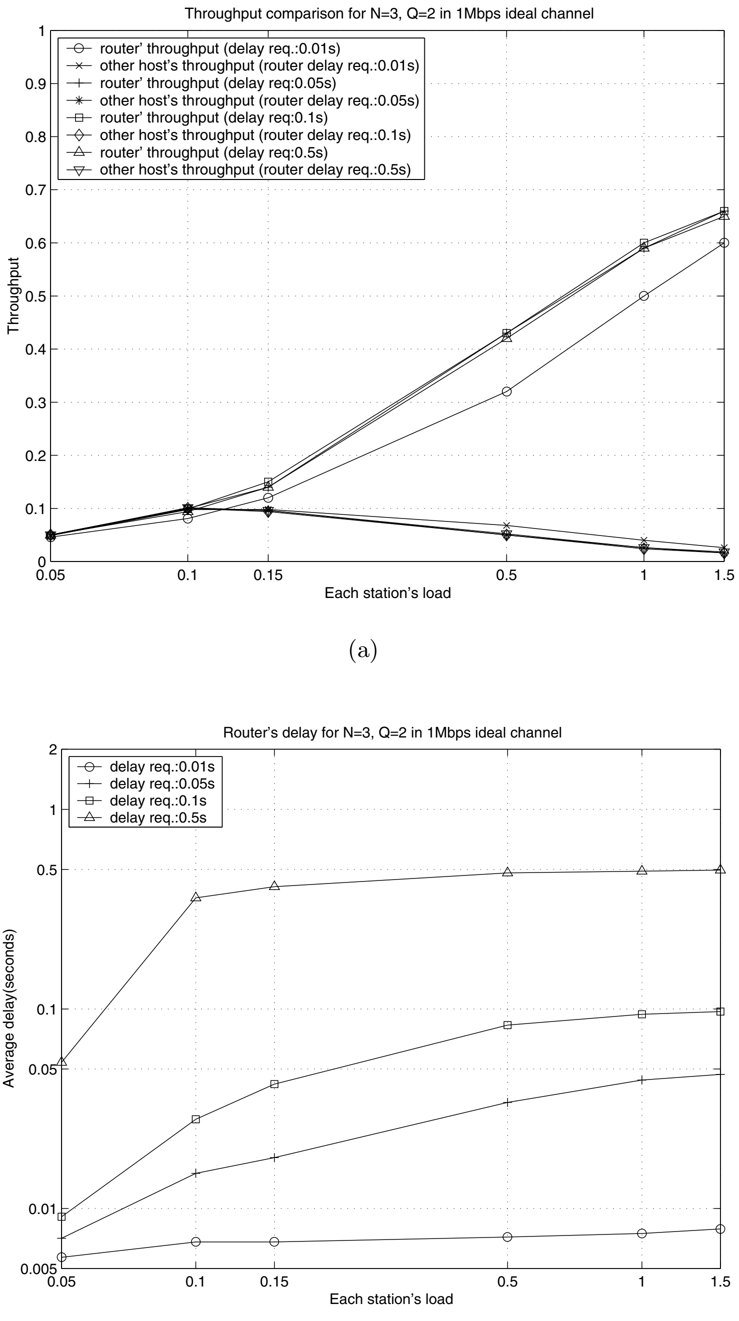


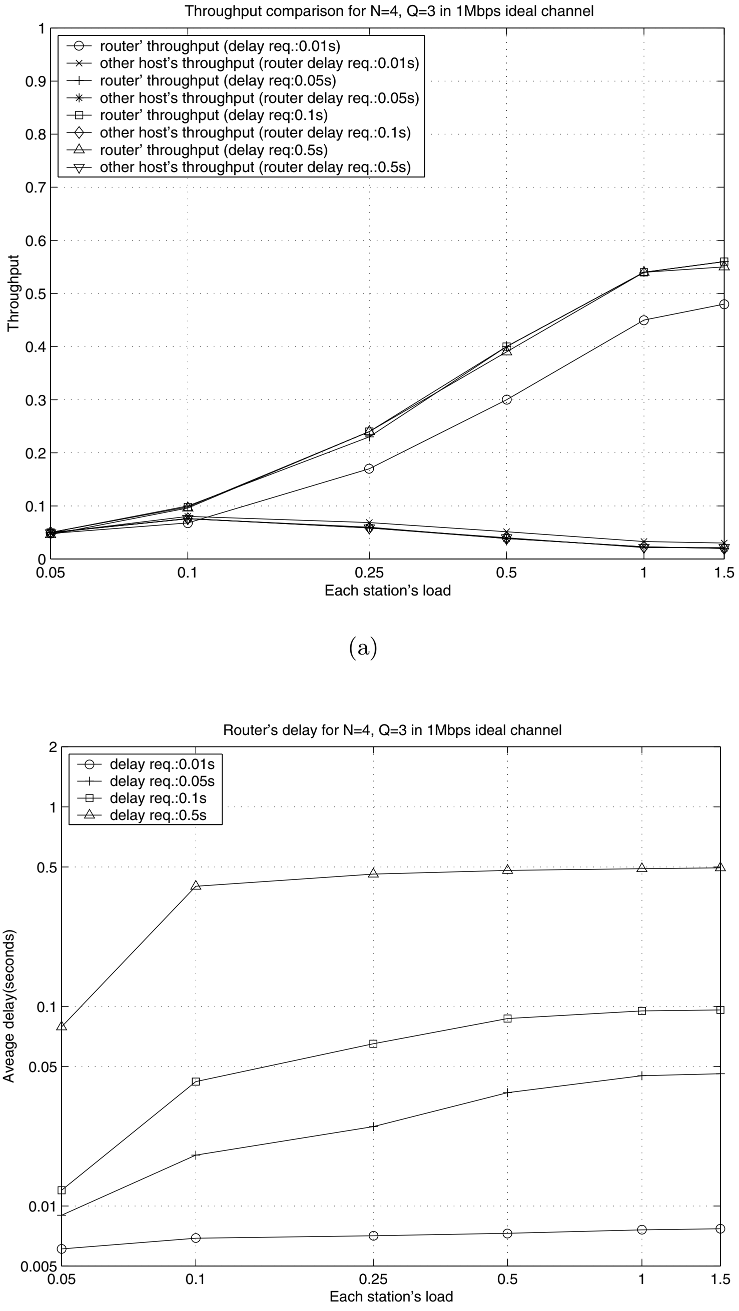
Fig. 5. hroughput v r u no gr



(a)

(b)

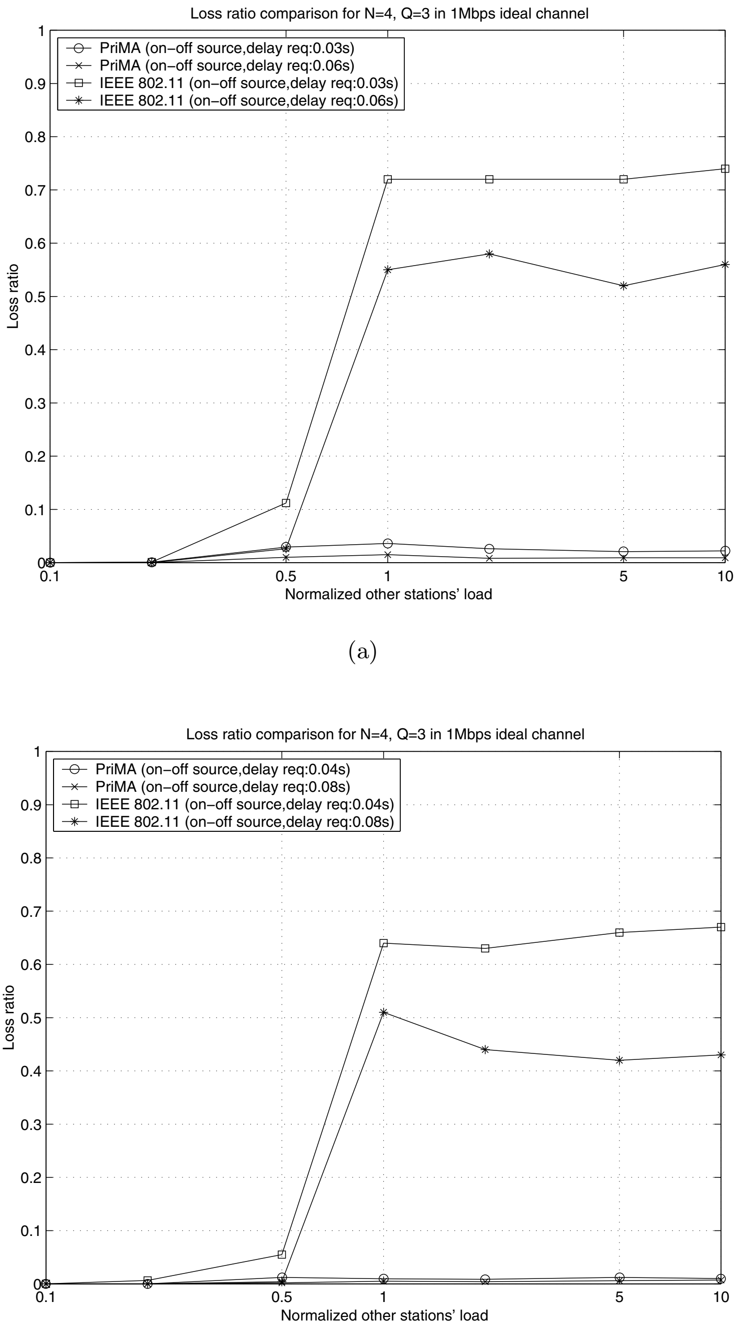
Fig. 6. On rout r on gur tion (3)



(a)

(b)

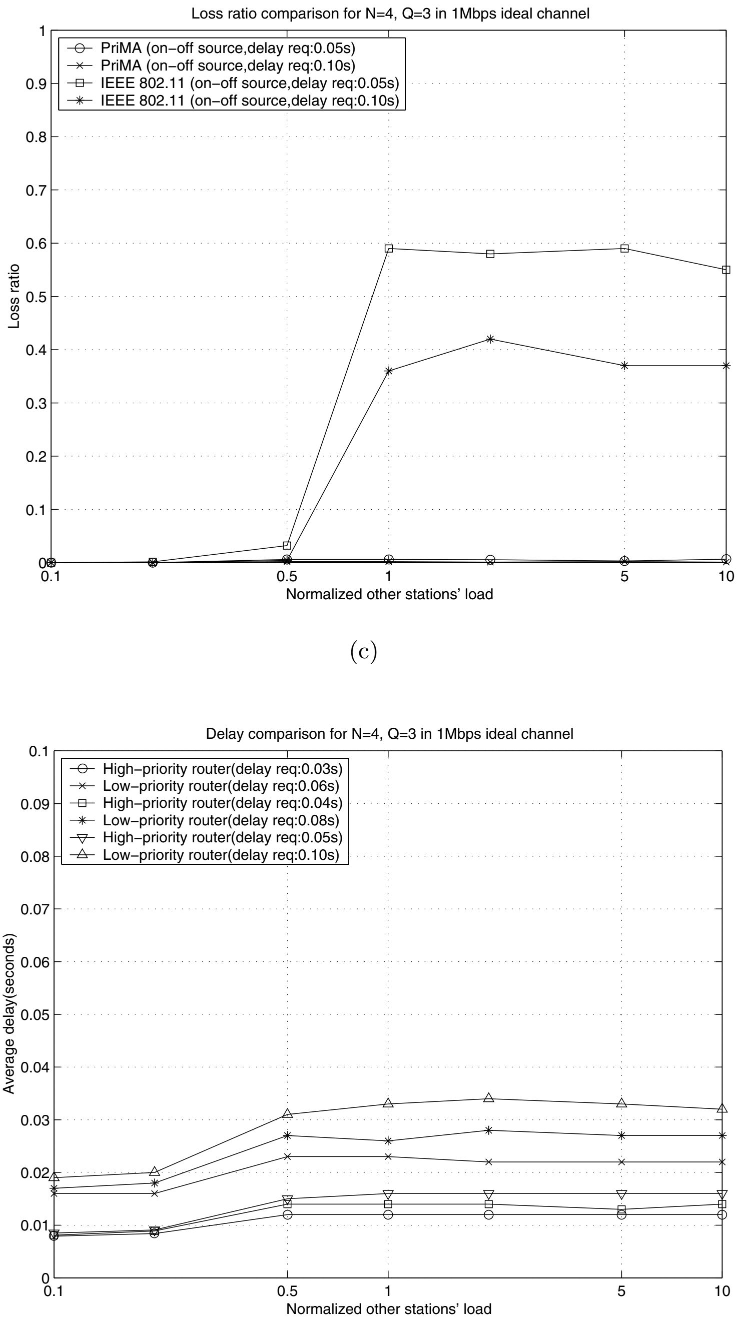
Fig. 7. On rout r on gur tion (4)



(a)

(b)

Fig. 8. o iff r nti tion



(c)

(d)

Fig. 8. o iff r nti tion (on t)

ho routing om p k t n giv n high r prioriti n g t liv r rli r th n oth r . imul tion r ult how th t riM i n p i ll promi ing M proto ol or ho n twork in pr n o h t rog n ou tr n o r quir m nt .

References

[1] http://www.ietf.org/html.charters/manet-charter.html.

[2] F. A. Tobagi and L. Kleinrock, "Packet Switching in Radio Channels: Part II - the Hidden Terminal Problem in Carrier Sense Multiple-access Modes and the Busy-tone Solution," in *IEEE Transactions on Communications, vol. COM-23, no. 12*, pp. 1417–1433, 1975.

[3] P. Karn, "MACA - a New Channel Access Method for Packet Radio," in *ARRL/CRRL Amateur Radio 9th Computer Networking Conference*, pp. 134–140, ARRL, 1990.

[4] J. Garcia-Luna-Aceves and C. L. Fullmer, "Performance of Floor Acquisition Multiple Access in Ad-Hoc Networks," in *Proceedings of 3rd IEEE ISCC*, 1998.

[5] IEEE Computer Society LAN MAN Standards Committee, ed., *IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications.* IEEE Std 802.11-1997, The Institute of Electrical and Electronics Engineers, New York, 1997.

[6] H. S. Chhaya and S. Gupta, "Throughput and Fairness Properties of Asynchronous Data Transfer Methods in the IEEE 802.11 MAC Protocol," in *6th International Conference on Personal, Indoor and Mobile Radio Communications*, 1995.

[7] J. L. Sobrinho and A. S. Krishnakumar, "Quality-of-Service in Ad Hoc Carrier Sense Multiple Access Wireless Networks," in *IEEE Journal on Selected Areas in Communications, Vol. 17, No. 8*, pp. 1353–1368, 1999.

[8] A. Muir and J. J. Garcia-Luna-Aceves, "Supporting real-time multimedia traffic in a wireless LAN," in *Proc. SPIE Multimedia Computing and Networking*, pp. 41–54, 1997.

[9] M. L. Jiang, J. Y. Li, and Y. C. Tay, "Cluster Based Routing Protocol(CBRP) Functional Specification." draft-ietf-manet-cbrp-spec-00.txt, Aug. 1998. Work in progress.

[10] R. Sivakumar, P. Sinha, and V. Bharghavan, "Core Extraction Distributed Ad hoc Routing (CEDAR) Specification." draft-ietf-manet-cedar-spec-00.txt, Oct. 1998. Work in progress.

[11] C. L. Fullmer and J. J. Garcia-Luna-Aceves, "Solutions to Hidden Terminal Problems in Wireless Networks," in *Proceedings of ACM SIGCOMM*, 1997.

Survivability Analysis of Ad Hoc Wireless Network Architecture

Krishna Paul[1] , Romit RoyChoudhuri[2], Somprakash Bandyopadhyay[3]

[1] Cognizant Technology Solutions, Sector V, Saltlake
Calcutta 700 091 India
{PKrishna2@cal.cts-corp.com}

[2] Department of Computer Sc and Engg
Haldia Institute of Technology
Haldia, West Bengal, India

[3] PricewaterhouseCoopers, Saltlake Technology Center
Sector V, Calcutta 700 091, India
{somprakash.bandyopadhyay@in.pwcglobal.com}

Abstract. Mobile ad hoc wireless networks are generating novel interests in mobile computing. The dynamism in network topology has thrown up multifarious issues, requiring a fresh look into the aspects of system design and networking protocols. As a direct consequence of injecting mobility into a static network, the formal relationships between several governing parameters have undergone changes. In this paper we have assayed the behavior of the ad hoc network as a whole and analyzed trends in the inter-parameter dependencies, with the objective of addressing to the survivability issues. We have finally drawn out an operating region of survivability for mobile ad hoc wireless networks in terms of user declared specifications. Our own simulator has been operative through the work. We have derived our survivability constraints from several runs of the network simulator.

1. Introduction

An ad hoc network [1] can be envisioned as a collection of mobile routers, each equipped with a wireless transceiver, which are free to move about arbitrarily. The mobility of the routers and the variability of other connecting factors results in a network with a potentially rapid and unpredictable changing topology. These networks may or may not be connected with the infrastructure such as internet, but still be available for use by a group of wireless mobile hosts that operates without any base-station or any centralized control. Applications of ad hoc networks include military tactical communication, emergency relief operations, commercial and educational use in remote areas, etc. where the networking is mission-oriented and / or community-based.

There has been a growing interest in ad hoc networks in recent years [1,2,3,4,5]. The basic assumption in an ad-hoc network is that, two nodes willing to communicate may be outside the wireless transmission range of each other but still be able to

C.G. Omidyar (Ed.): MWCN 2000, LNCS 1818, pp. 31-46, 2000.

communicate if other nodes in the network are willing to forward packets from them. However, the successful operation of an ad-hoc network will be hampered, if an intermediate node, participating in a communication between two nodes, moves out of range suddenly or switches itself off in between message transfer. The situation is worse, if there is no other path between those two nodes. An important problem associated with this is to find a stable path satisfying multiple constraints to ensure certain level of QoS guarantee during communication.

Lot of research has been done on ad hoc network routing protocols in order to solve the problem of routing packets. However, there is no complete proposal available to assess the survivability issues in ad hoc network in order to provide a network specification to support effective communication in such a dynamic environment. Survivability analysis [6], in this context, can be defined as network specifications and management procedures to minimize the impact of system dynamics on the network services . For example, assume an area 1000 x 1000 sq.meter where 20 nodes are moving around with an average velocity of 10m/sec. The transmission range for each node is, say, 350 meter. Under a given traffic pattern, will this network be able to provide the required service guarantee to its users in spite of the dynamic change in topology due to mobility? If the answer is yes, let us further assume that some of the users decide to switch-off or to leave the field or to increase their mobility. Will the network still survive? On the other extreme, let us assume that 20 new nodes join the system, making the node count to 40. The transmission range that is optimal for 20 nodes may be too high for 40 nodes, as this will increase collision and congestion of control / data packets. Will the network still be able to provide the required service guarantee to its users ?

So, survivability analysis and drawing up a specification for a survivable ad hoc network is an important issue that we want to address in this paper.

2. Survivable Systems

2.1 Definition and Characteristics of Survivable Systems

Traditionally, survivability in network systems has been defined as the capacity of a system to fulfil its mission, in a timely manner, in the presence of failures [7]. The term mission refers to a set of very high-level requirements or goals. Timeliness is a critical factor that is typically associated with the mission fulfillment. In the context of ad hoc network, mission fulfillment in a timely manner implies that the network should be able to ensure certain level of service guarantee to its user in the presence of system dynamics. Survivability analysis, in this context, can be defined as network specifications and management procedures to minimize the impact of system dynamics on the network services.

Thus, in this study, we are not considering failure due to hardware malfunctions or software errors. However, in an ad hoc network, any node can randomly switches itself off causing an event equivalent to node failure. Similarly, any link between two node can get disconnected anytime because of mobility of the nodes, causing an event equivalent to link failure. Additionally, new nodes can join the system at any point of time; similarly, new links can be formed between any two nodes, as they come closer to each other due to their mobility.

For survivability, we must achieve system-wide properties that typically do not exist in individual nodes. A survivable system must ensure that desired survivability properties emerge from interactions among the components in the construction of reliable systems from unreliable components [7]. If survivability properties are emergent they are present only when the number of nodes of a system are sufficiently large. If the number or arrangements of nodes falls below a critical threshold, the attendant survivability property fails. For example, we can specify an ad hoc network to operate at a transmission range of, say, 350 with number of nodes between 15 and 20 at a mobility ranging from 5m/s to 20m/s. But, if number of nodes falls below that, the system may not survive i.e. may not be able to ensure certain service guarantee to its users.

2.2 Specifying the Requirements of Survivable Ad Hoc Network

Central to the notion of survivability analysis is to identify and ensure the maintenance of certain essential attributes and the operating levels of those attributes that must be associated with the specified level of service guarantee. In the context of ad hoc network , the goal is to maintain the network availability and allow the data packets to be delivered to the intended destination from a source in spite of the changes in network topology due to its dynamic behavior. Survivability analysis consists of determining whether service objectives can be maintained during all operational modes.

Thus, network service in the context of ad hoc network is primarily pivotal to two fundamental requirements:

1. establishing a connection between any two nodes in the network at any instant of time.
2. Assuring an uninterrupted connection until a finite volume of data transfer has been accomplished (of course with limited delay in data transfer).

Survivability issues depend entirely on how well these two demands are met with. A network would be called survivable if it meets both the above requirements satisfactorily. Now, in order to declare an ad hoc network survivable we need to first define the user requirements in more formal terms. In other words we require a set of metrics which would inherently take care of all the service demands and finally throw up a numerical values depicting the degree of survivability for a given set of design specifications. Our objective is to design such a set of metrics in terms of the basic network parameters: number of nodes (N), transmission range (R), mobility (M), average volume of data to be communicated from a source to its destination (V) and average number of communication events per minute (C).

3. System Description

The network is modeled as a graph G = (N,L) where N is a finite set of nodes and L is a finite set of unidirectional links. Each node $n \in N$ is having a unique node identifier. Since in a wireless environment, transmission between two nodes does not necessarily work equally well in both direction [1], we assume unidirectional links. Thus, two nodes n and m are connected by two unidirectional links $l_{nm} \in L$ and $l_{mn} \in$ L such that n can send message to m via l_{nm} and m can send message to n via l_{mn}. However, in this study, we have assumed $l_{nm} = l_{mn}$ for simplicity.

In a wireless environment, each node n has a wireless transmitter range. We define the neighbors of n, $N_n \in N$, to be the set of nodes within the transmission range R of n. It is assumed that when node n transmit a packet, it is broadcast to all of its neighbors in the set N_n. However, in the wireless environment, the strength of connection of all the members of N_n with respect to n are not uniform. For example, a node $m \in N_n$ in the periphery of the transmission range of n is weakly connected to n compared to a node $p \in N_n$ which is more closer to n. Thus, the chance of m going out of the transmission range of n due to outward mobility of either m or n is more than that of p.

Each link l_{nm} is associated with a signal strength S_{nm} which is a measurable indicator of the strength of connection from n to m at any instant of time. Due to the mobility of the nodes, the signal strengths associated with the links changes with time. When the signal strength S_{nm} associated with l_{nm} goes below a certain threshold S_t, we assume that the link l_{nm} is disconnected.

Affinity a_{nm}, associated with a link l_{nm}, is a prediction about the span of life of the link l_{nm} in a particular context [5]. For simplicity, we assume a_{nm} to be equal to a_{mn} and the transmission range R for all the nodes are equal. To find out the affinity a_{nm} , node n sends a periodic beacon and node m samples the strength of signals received from node n periodically. Since the signal strength is roughly proportional to $1/R^2$, we can predict the current distance d at time t between n and m. If M is the average velocity of the nodes, the worst-case affinity a_{nm} at time t is (R-d)/M, assuming that at time t, the node m has started moving outwards with an average velocity M. For example, If the transmission range is 300 meters, the average velocity is 10m/sec and current distance between n and m is 100 meters, the life-span of connectivity between n and m (worst-case) is 20 seconds, assuming that the node m is moving away from n in a direction obtained by joining n and m..

Given any path p = (i, j, k, …, l, m), the **stability of path p** [5] at a given instant of time will be determined by the lowest-affinity link (since that is the bottleneck for the path) and is defined as $\min[a_{ij}, a_{jk}, \ldots, a_{lm}]$. In other words, stability of path p between source s and destination d, $\eta^p_{sd,}$ is given by

$$\eta^p_{sd} = \min_{\forall i,j} a^p_{ij}$$

However, the notion of stability of a path is dynamic and context-sensitive. As indicated earlier, stability of a path is the span of life of that path from a given instant of time. But stability has to be seen in the context of providing a service. A path between a source and destination would be stable if its span of life is sufficient to complete a required volume of data transfer from source to destination. Hence, a given path may be sufficiently stable to transfer a small volume of data between source and destination; but the same path may be unstable in a context where a large volume of data needs to be transferred.

4. Route Discovery and Data Communication Mechanism in Ad Hoc Network

The existing routing protocol can be classified either as proactive or as reactive [3]. In proactive protocols, the routing information within the network is always known beforehand through continuous route updates. The family of distance vector and link state protocols is examples of proactive scheme. Reactive protocols, on the other hand, invoke a route discovery procedure on demand only. The family of classical flooding algorithms belongs to this group. It has been pointed out that proactive protocols are not suitable for highly mobile ad hoc network, since they consume large portion of network capacity for continuously updating route information. On the other hand, on-demand search procedure in reactive protocols generate large volume of control traffic and the actual data transmission is delayed until the route is determined.

Whatever may be the routing scheme, frequent interruption in a selected route would degrade the performance in terms of quality of service. In [5], we have attempted to minimize route maintenance by selecting stable routes, rather than shortest route, which is illustrated below.

4.1 Path Finding Mechanism

A source initiates a route discovery request when it needs to send data to a destination. The source broadcast a route request packet to all neighboring nodes. Each route request packet contains source id, destination id, a request id, a route record to accumulate the sequence of hops through which the request is propagated during the route discovery, and a count max_hop which is decremented at each hop as it propagates. When max_hop=0, the search process terminates. The count max_hop thus limits the number of intermediate nodes (hop-count) in a path.

When any node receives a route request packet, it decrements max-hop by 1 and performs the following steps:

1. If the node is the destination node, a route reply packet is returned to the source along the selected route, as given in the route record which now contains the complete path information between source and destination.
2. Otherwise, if max_hop=0, discard the route request packet.
3. Otherwise, if this node id is already listed in the route record in the request, discard the route request packet (to avoid looping).
4. Otherwise, append the node id to the route record in the route request packet and re-broadcast the request.

When any node receives a route reply packet, it performs the following steps:

1. If the node is the source node, it records the path to destination.
2. If it is an intermediate node, it appends the value of affinity and propagates to the next node listed in the route record to reach the source node.

4.2 Sending the Data from Source to Destination

When a source initiates a route discovery request, it waits for the route reply until time-out. If it receives a path, it computes its stability $\eta^{p}_{sd.}$ If V_{sd} is the volume of data to be send to destination and if B is the bandwidth for transmitting data, V_{sd} / B is the one-hop delay to transmit the data, ignoring all other delay factors. If H is the number of hops from source to destination, H* V_{sd} / B will be the time taken to complete the data transfer. If η^{p}_{sd} is sufficient to carry this data, the path is selected. Otherwise, the source checks the next path, if available, for sufficient stability. In order to check the sufficiency, η^{p}_{sd} is multiplied with a correction factor f, to be decided dynamically, to take care of estimation error and other delay factors related to traffic characteristics.

```
The Algorithm:
Step I: p:= 0;
Step II: **wait** for a path **until** timeout;
Step III: **if** a path is available **then**
**begin**
        p:=p+1;
        find $\eta^{p}_{sd} = \min_{\forall i,j} \eta^{p}_{ij}$ ; { find the stability of path k}
        **if** $(H^* V_{sd} / B) < f * \eta^{p}_{sd}$ {if the path is suffiently stable }
           **then** start sending $V_{sd}$ into $p_{th}$ path
        **else** reject the path and goto step II
**end**
Step IV: terminate.
```

5. The Simulation Environment

Existing simulators are not well-equipped to serve our purpose [9,10,11]. Hence, in order to model and study the survivability issues of the proposed framework in the context of ad hoc wireless networks, we have developed a simulator with the capability to model and study the following characteristics:

- Node mobility
- Link stability (*affinity*)
- Affinity- based path search
- Dynamic network topology depending on number of nodes, mobility and transmission range
- Realistic physical and data link layers in wireless environment
- Data communication with different data volume and different frequency of communication events per minute.

The proposed system is evaluated on a simulated environment under a variety of conditions. In the simulation, the environment is assumed to be a closed area of 1000 x 1000 sq. meter in which mobile nodes are distributed randomly. We ran simulations for networks with different number of mobile hosts, operating at different transmission ranges. The bandwidth for transmitting data is assumed to be 1000 packets / sec. The packet size is dependent on the actual bandwidth of the system.

In order to study the delay, throughput and other time-related parameters, every simulated action is associated with a simulated clock. The clock period (time-tick) is assumed to be one millisecond (simulated). For example, if the bandwidth is assumed to be 1000 packets per second and the volume of data to be transmitted from one node to its neighbor is 100 packets, it will be assumed that 100 time-ticks (100 millisecond) would be required to complete the task. The size of both control and data packets are same and one packet per time-tick will be transmitted from a source to its neighbors.

The speed of movement of individual node ranges from 5 m/sec. to 20 m/sec. Each node starts from a home location, selects a random location as its destination and moves with a uniform, predetermined velocity towards the destination. Once it reaches the destination, it waits there for a pre-specified amount of time, selects randomly another location and moves towards that. However, in the present study, we have assumed zero waiting time to analyze worst-case scenario.

6. Analyzing the Impact of Dynamic Topology on Survivability

6.1 Related Definitions

To conceive certain trends in network characteristics on the whole, some terms have been used that are defined as follows:

Average Connectivity Efficiency (E): Connectivity Efficiency has been defined as the ratio of total number of connected node-pairs (in single hop or in multiple hops) and the total number of available node pairs at any instant of time. This fraction captures the degree of connectivity among the nodes in any snapshot of the mobile environment. From the survivability point of view, this parameter is an indicator to the success rate of a source node, in attempting to establish a connection with a destination node. The efficiency values obtained over several snapshots (taken at intervals of one second from the simulator) of the dynamic environment have been finally averaged to yield the Average Connectivity Efficiency. A network where all the node-pairs are always connected in single or multiple hops have a Average Connectivity Efficiency of 100%. Thus,

$$\textbf{Average Connectivity Efficiency (\%)} = \frac{\Sigma^{T}_{i=1}\textbf{ (no. of connected node pairs) * 100}}{\textbf{T * Number of node-pairs}}$$

Average Network Stability (S): From survivability perspectives, the span of time for which two nodes remain connected (given the number of nodes, transmission range and the mobility) need to be analyzed. A parameter, **affinity**, introduced in [5] and explained in section 3 has been used for average worst case analysis. As explained in section 3, the stability of the path (i.e. the span of time for which this path would exist) can be determined by the weakest link in the path.

Two nodes in the ad hoc environment may often be connected with several paths. For data communication between two nodes, the best path should always be chosen i.e. the path assuring greater stability. Thus,
Node to Node Stability = *max* (stability of all the paths between the two nodes).

The **Average Network Stability** has been defined as the average node to node stability over time.

$$\textbf{Average Network Stability} = \frac{\Sigma^{T}_{i=1}\,\Sigma_{\text{all node-pair}}\,\textbf{(Node to Node Stability)}}{\textbf{T * number of node-pairs}}$$

Average Number of Neighbors (G): The study of percolation is an important aspect from the data communication point of view in a mobile computing environment [12]. For a random distribution of nodes in a bounded region, percolation is proportional to the number of neighbors, which in turn is a function of node density and signal strength. Average Number of neighbors has been defined as:

$$\textbf{Average Number of Neighbors} = \frac{\Sigma^{T}_{i=1}\Sigma_{\text{all node}}\,\textbf{(Number of neighbors of each node)}}{\textbf{T * number of node}}$$

6.2 Variation of Average Connectivity Efficiency (E) with N, R, and M

It is quite obvious that if the signal strength increases, the probability of connectivity also increases. The variation of connectivity efficiency against signal strength has been shown in the plot in fig.1(a). However this signal strength cannot be allowed to increase indefinitely due to other overheads:

1. Cost (power consumption due to battery usage) increases as the signal strength is raised.
2. Congestion and collision are the inevitable outcome of higher signal strength during data communication, as will be illustrated in the next section.

A larger number of users in a closed area indicate a higher node density. Since E is a measure of connectivity and connectivity is heavily dependent on how close the nodes are with each other, the total number of nodes in a bounded area also contributes to the connectivity efficiency. Thus, the connectivity efficiency bears a composite relation with the number of nodes as well (Fig. 1(b)). From figure 1, it is quite evident that, to achieve a specific threshold of efficiency, there is a lower cut off of the signal strength for a given number of nodes.

E would not depend on the mobility of the nodes. If the node mobility is high, then the probability of nodes moving out of a node's transmission range increases as much as the probability of new nodes coming into the transmission range of the same node. As a result, average value of connectivity taken over a long time remains unaffected at different mobility.

Figure 2 depicts the variation of Average Connectivity Efficiency (E) against Average Number of Neighbors (G). Although G does not reflect the actual dependence of E on N and R, it can be instrumental in deciding the cut off for satisfactory connectivity in the network. Over G=6 the efficiency is always found to

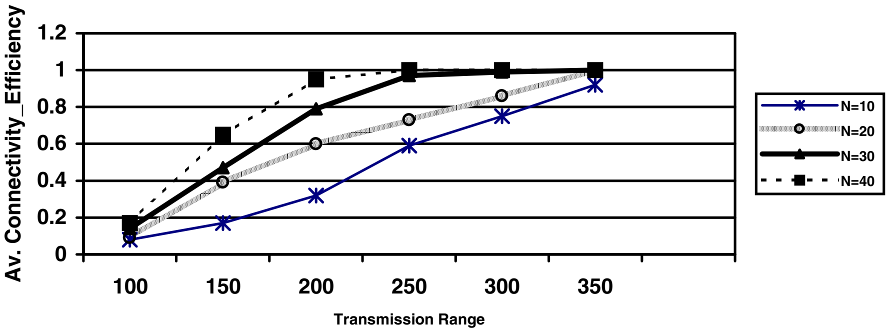


Fig 1(a). Average Connectivity Efficiency vs. Transmission Range for different number of nodes

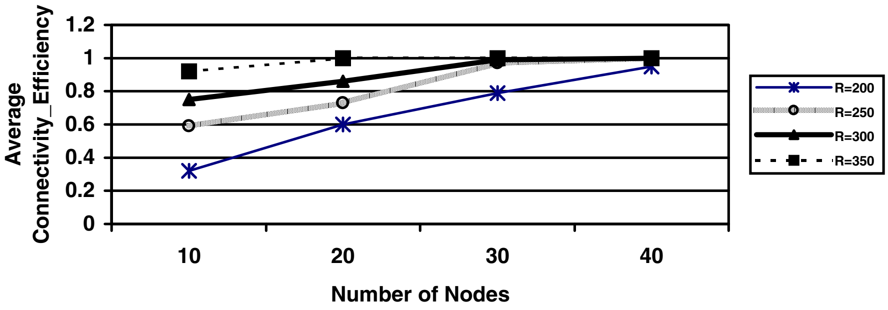


Fig 1(b). Average Connectivity Efficiency vs. Number of Nodes for different Transmission Range

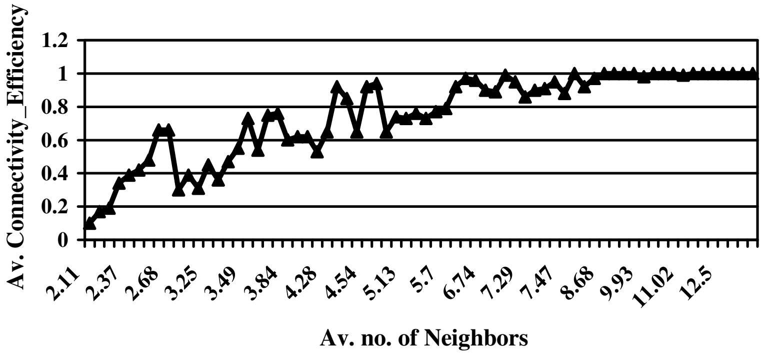


Fig 2. Average Connectivity Efficiency vs. Average Number of Neighbors

be increasing over 0.8. Over a certain threshold of neighbors, the network becomes connected and further increase in G would only hike overhead. The optimal value of G as six to eight has been proposed earlier[12,13] for large number of nodes which has been revalidated here with lower number of nodes with different mobility pattern.

Hence, the conclusion from above is: to ensure a fairly high level of connectivity, the design parameters should be such that the predicted number of neighbors is greater than 6. Assuming uniform distribution, the average node density per unit area is N/A, where N is the number of nodes in area A. Assuming uniform transmission range R, $[(N*\Pi R^2/A) - 1]$ will be the average number of neighbors, which should be greater than six to have E >0.8.

6.3 Variation of Average Network Stability against N. R, and M

From the perspective of network service, the stability of a path between two arbitrary nodes indicates the volume of data that could be transferred between the two nodes in question (provided none of the intermediate nodes switch off during data transfer). Conversely, it is stability, which would be instrumental in deciding the thresholds of average transferable data volume, thus ensuring survivability.

A high node density in the operating environment essentially indicates that the average distance between two nodes is less in comparison to a model of low node density. Naturally, if two nodes remain in greater proximity, for a given signal strength and mobility, they would remain in contact for a longer period of time. Consequently the average stability of links would be higher and thus the stability of paths. Thus, it can be said that the average stability (S) of a mobile ad hoc wireless network would increase with increase in node density or N (as node density = N / A). At the same time, if the average affinity of links in a network features to be high, the average stability of paths would also be correspondingly higher. From the expression of affinity, we see that affinity of a link increases with increase in transmission range and/or decrease in mobility. Stability can thus be said to be directly proportional to transmission range and inversely proportional to mobility (Figure 3).

7. Analyzing the Impact of Route Discovery and Data Communication on Survivability

The above analysis does not take into account the congestion and collision factors that would happen during data communication. We will show that even if a network is well-connected, it may not guarantee successful data communication.

7.1 Related Definitions

Route Discovery Efficiency is defined as the ratio of the average number of route replies obtained per minute and the average number of route request generated per minute. As discussed, the number of route request generated per minute would depend on the number of communication events initiated per minute (C). However,

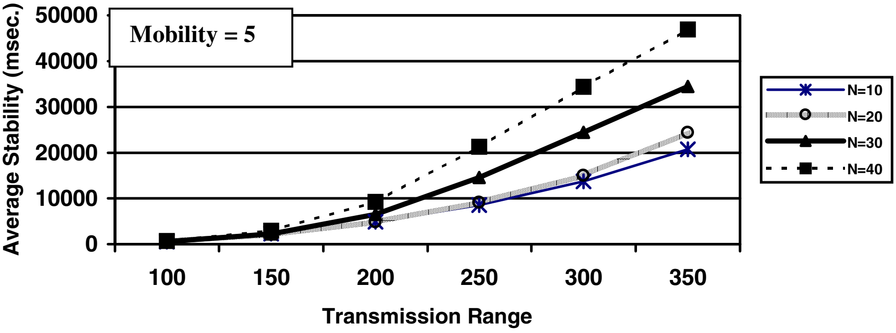


Fig 3(a). Average Network Stability vs. Transmission Range at mobility =5.

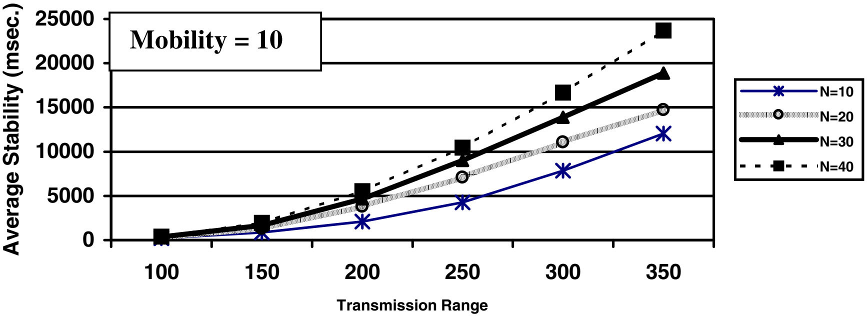


Fig 3(b). Average Network Stability vs. Transmission Range at Mobility = 10.

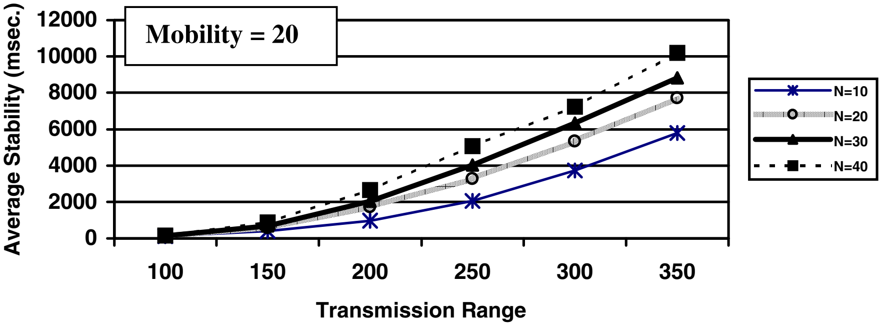


Fig 3(c). Average Network Stability vs. Transmission Range at Mobility =20.

the success of route request i.e. getting a rout reply back within a reasonable period of time (500 msec in our case) would depend on the degree of collision and congestion of the network. This is not only dependent on E but also on the average volume of data communicated from a source to its destination (V) and frequency of communication events per minute (C). If C and / or V increases, the probability of collision and congestion would increase, which in turn will affect the Roure Discovery Efficiency.

Service Efficiency is defined as the ratio of the average number of communication events successful within a reasonable period of time per minute and the average number of route request generated per minute. Service Efficiency depends on four factors : 1) Route request has been generated but route reply has not come back within a reasonable period of time, 2) Route replies have been obtained but the paths are rejected because they are not stable enough to carry out the required volume of data transfer, 3) A path is selected and data communication has started but the path could not be retained throughout the entire period of data communication, and 4) the network delay is too high to complete the data transfer within a reasonable period of time. It has been shown in [5] that the use of stability based routing reduces the probability of (3). However, the prediction of stability would be affected, if the network is heavily congested which will in turn affects the Service Efficiency.

7.2 Variation of Route Discovery Efficiency against N,R, and V with C=4 per Minute

For a given number of nodes, the number of control packets generated increases drastically beyond a certain transmission range. In a collision-free environment, if G is the average number of neighbors and max_hop =4, then the number of control packet generated will be G^4 per communication event. Therefore, it is obvious that with increase in G, the number of control packets increases drastically.

The congestion due to control packets at high transmission range would affect the Route Discovery Efficiency as shown in Figure 4. The effect would be more pronounced for larger number of nodes and for larger amount of data volume. Here, number of communication event per minute (C) is assumed to be 4. We have also studied this variation with C=10 (not shown) where the large data volume would degrade the Route Discovery Efficiency further. In any case, for a fixed number of N, there is an optimum value of R, R^{Nopt}, which will maximize the route discovery efficiency. Increasing R beyond that point will degrade the performance.

However, R^{Nopt} alone can not maximize route discovery efficiency. We need to consider two more factors : average volume of data to be communicated from a source to its destination (V) and average number of communication events per minute (C). The system should be capable of absorbing the control and data packets before a new communication event starts.

Depending on R^{Nopt} and the average network stability at that R, we can specify V for an average mobility M. If we increase M or V beyond that, the Service Efficiency will suffer.

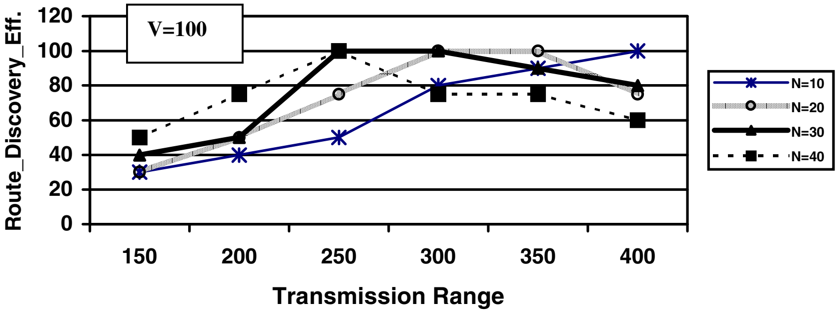


Fig 4(a). Route_Discovery_Efficiency vs. Transmission Range with ***Data Volume = 100 packets, Max_Hop=4*** **and** ***No of Communication = 4 / min.***

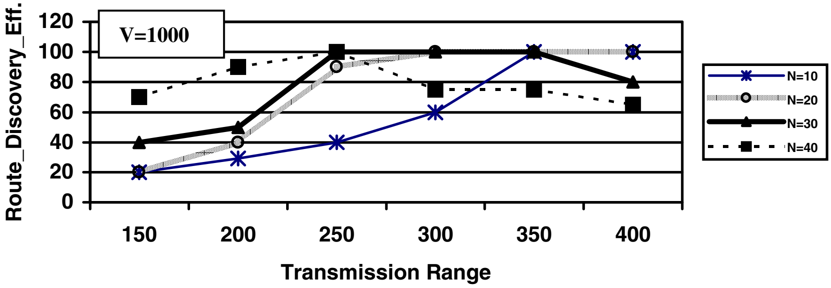


Fig 4(b). Route_Discovery_Efficiency vs. Transmission Range with ***Data volume= 1000 packets, Max_Hop=4*** **and** ***No of Communication = 4 / min.***

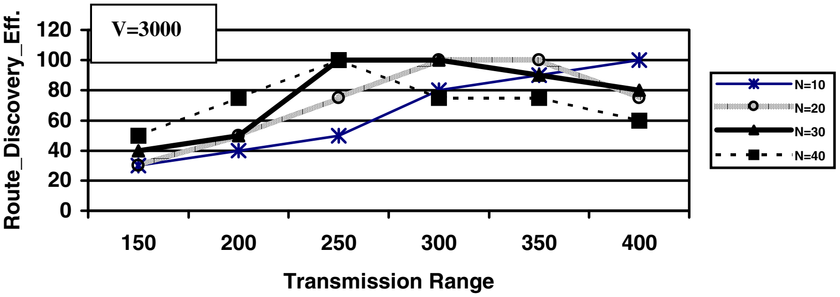


Fig 4(c). Route_Discovery_Efficiency vs. Transmission Range with ***Data volume =3000 packets, Max_Hop=4 and No of Communication = 4 / min.***

7.3 Variation of Service Efficiency with C=4 per Minute

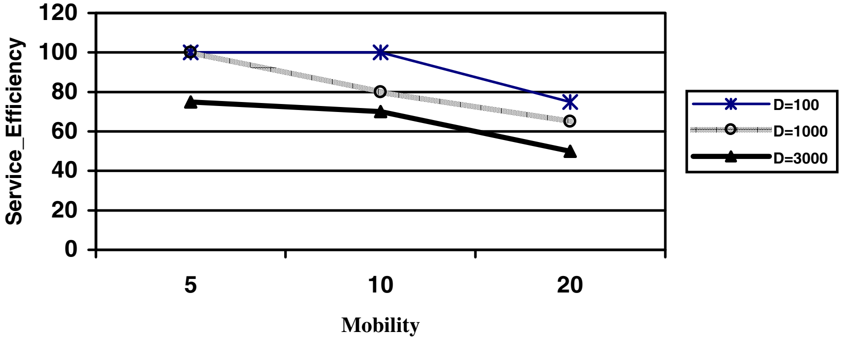
For a given number of node and corresponding R^{Nopt} , the variation of Service Efficiency against M and V is shown in figure 5. It is evident that getting a high Service Efficiency with V=3000 is difficult to obtained in this set up where mobility is varying between 5 to 20. The reason is that we are not getting sufficient stable paths to complete the data transfer. On the other hand, for lower volume of data and low mobility, it is possible to get a Service Efficiency > 80%. With M=20, getting a high Service Efficiency for a data volume > 1000 is difficult, when the number of nodes are more than 20.

8. Survivability Metrics and Specifications for a Survivable Ad Hoc Network

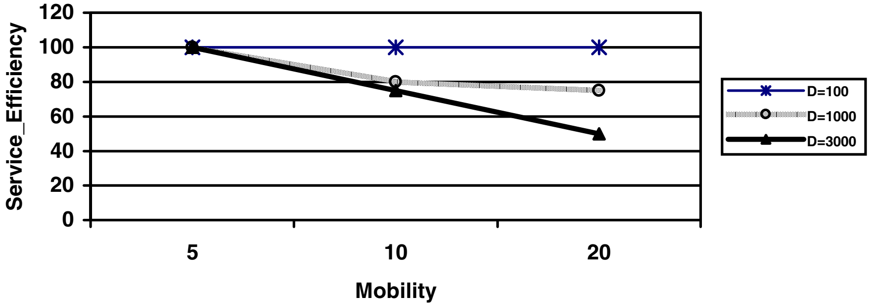
From the above analyses, the following points can be concluded :

- For a fixed number of N, the Average Connectivity Efficiency will be more than 0.8 beyond a certain value of R. If we increase R further, the Connectivity Efficiency will improve and saturate to 1.0. Consequently, the Average Stability will also improve so that a larger volume of data could be sent. But the Route Discovery Efficiency and, consequently, the Service Efficiency will go down because of large number of control packets and / or data packets.
- For a fixed number of N, there is an optimum value of R, R^{Nopt}, which will maximize the route discovery efficiency.
- However, R^{Nopt} alone can not maximize route discovery efficiency. We need to consider two more factors : average volume of data to be communicated from a source to its destination (V) and average number of communication events per minute (C). The system should be capable of absorbing the control and data packets before a new communication event starts.
- Depending on R^{Nopt} and the average mobility M, we can specify average network stability which will in turn determine V. If we increase M or V beyond that, the Service Efficiency will suffer.

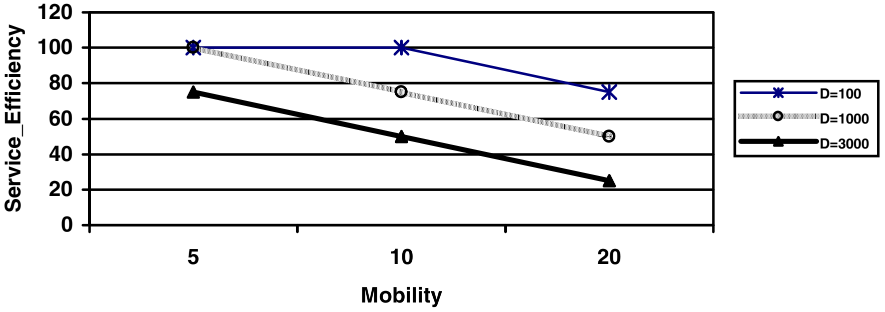
The aim of our entire analysis is to model the survivability region of operation for a mobile ad hoc wireless network. In other words, we need to answer questions like :
What should be the transmission range of operation and the maximum mobility for an ad hoc network with 30 users, if the user require a Service Efficiency of 80% and 1000-Kb average data volume for transfer? The kind of answers we are trying to provide is that, for 30 users with transmission range between 275 m to 325 m, it is possible to achieve the required Service Efficiency with V<=1000 and C=4, if the average mobility is less than 10. As another example, suppose we ask that: what is the Service Efficiency achievable if the number of users are 35 to 40, moving with an average velocity between 10 to 20 m/sec and the average data transfer requirement is 4 per minute with an average volume of data = 100 packets ? From the above analyses, we can say that with R=250, we can achieve a Service Efficiency of around 80%.



5(a). Service_Efficiency vs. Mobility with No. of Comm.=4/min. and *N=20* & *R=350* .



5(b). Service_Efficiency vs. Mobility with No. of Comm.=4/min and *N=30* & *R=300*



5(c). Service_Efficiency vs. Mobility with No. of Comm.=4 / min and *N=40* & *R=250* .

9. Conclusion

In this analysis, we have not included the impact of the variation of C. We have also not included the per-hop delay and delivery delay under different conditions. However, this preliminary analysis illustrates the basic interdependencies among several governing parameters that would help us in drawing up specifications for survivable ad hoc networks.

Reference

[1] D. B. Johnson and D. Maltz, Dynamic source routing in ad hoc wireless networks, T. Imielinski and H. Korth, eds., *Mobile computing,* Kluwer Academic Publ. 1996.

[2] S. Corson, J. Macker and S. Batsell, Architectural considerations for mobile mesh networking, Internet Draft RFC Version 2, May 1996.

[3] Z.J.Haas, A new routing protocol for the reconfigurable wireless networks, ICUPC'97, San Diego, CA, Oct. 1997.

[4] V. D. Park and M. S. Corson, A highly adaptive distributed routing algorithm for mobile wireless networks, Proc. IEEE INFOCOM '97, Kobe, Japan, April 1997.

[5] K. Paul, S. Bandyopadhyay, D. Saha and A. Mukherjee, Communication-Aware Mobile Hosts in Ad-hoc Wireless Network, Proc. of the IEEE International Conference on Personal Wireless Communication, Jaipur, India, Feb. 1999.

[6] David Tipper, Sreeniwas Ramaswami and Teresa Dahlberg, PCS Networks Survivability, to appear in IEEE WCNC 99, New Orleans, LA, USA

[7] R.J.Ellison D.A. Fisher R.C. Linger H. F. Lipson T. Lonstaff, N. R. Mead, Survivable Network System: An Emerging Discipline, Technical Report CMU/SEI-97-TR-013, Carnegie Mellon University, November, 1997.

[8] D. Medhi, A Unified Approach to Network Survivability for Teletraffic Networks: Models, Algorithms and Analysis, IEEE Transactions on Communications April 1994.

[9] K. Fall, and K. Varadhan. ns Notes and Documentation. The VINT Project, UC Berkeley. http://www-mash.cs.berkeley.edu/ns/,1997.

[10] J. Short, R. Bagrodia and L. Kleinrock. „Mobile Wireless Network System Simulation." *Wireless Network Journal 1*, no. 4, 1995.

[11] Josh Broch,; D.A. Maltz; D.B. Johnson; Y. Hu and J. Jetcheva. „A Performance Comparison of Multi-Hop Wireless Ad Hoc Network Routing Protocols." In *Proceedings of the Fourth Annual ACM/IEEE International Conference on Mobile Computing and Networking* (Mobicom'98), Dallas, Texas, Oct. 25-30, 1998.

[12] Y.C.Cheng and T.G.Robertazzi, Critical connectivity phenomena in multihop radio network, IEEE Trans. Commun. , 37(1989), pp 770-777.

[13] H.Takagi and L.Kleinrock, Optimal transmission ranges for randomly distributed packet radio terminals, IEEE Trans. Commun. vol COM-32, pp246-257, Mar.1984.

A Token Passing Tree MAC Scheme for Wireless Ad Hoc Networks to Support Real-Time Traffic

o nq ng[1] ng h ngm ng[2] n j ng[1]

[1] Department of Electrical Engineering, National University of Singapore,
10 Kent Ridge Crescents, Singapore, 119260
{engp9040, engp8587}@nus.edu.sg
[2] Centre for Wireless Communications, National University of Singapore,
20 Science Park Road, #02-34/37, Singapore, 117674
jiangsm@cwc.nus.edu.sg

Abstract. This paper introduces a distributed MAC protocol which can provide delay-bounded service in wireless ad hoc networks. Ad hoc networks are network architectures that do not rely on a pre-existing fixed infrastructure, which imposes heavy challenges on designing MAC scheme with QoS support. Most existing MAC protocols for wireless ad hoc networks have been designed to support none-delay-sensitive applications. Given an increasing demand on supporting multimedia applications, there is much interest in designing MAC schemes to satisfy the requirements of real-time applications in ad hoc environments. The timed token based MAC scheme proposed in this paper allocates the bandwidth to the users according to their requirements. This scheme can guarantee the deadline of real-time traffic even in the case of heavy traffic load. In addition, a logical token-passing tree structure adopted in this scheme overcomes the hidden terminal problems in ad hoc networks.

1 Introduction

n h r s non o pr x s ng n r s ru ur w r l ss ho n work n r p ly ploy . hus su h n works wh h n op r n r n n work on ons prov low os n fl x l solu on o ommun on n works. Mos n r s n h s hnology om s rom o h h omm r l m rk n m l ry ppl ons. h r r m ny h ll ng s o h urr n r s r h n h s r . On o hos ssu s s m um ss on rol (M) wh h on rols no s o ss m um n rm n s qu l y o s rv (o) n h nn l u l z on 1 . Mos x s ng M pro o ols wh h n prov o p l y o us rs r qu r som n r l n s .g. h s -s ons n llul r w r l ss n works. u h s h m s r no su l n ho nv ronm n s us o h l k o n r l n s s m n on ov . l hough no n s l o un on s n r l un mpor r ly h s no nno p r orm s ro us ly s x on us o h poss l y o pl orm mov m n s. h r or h s r u M s h m s r mor s r l h n h n r l z M s h m s us n llul r sys ms l hough h l r suppor r l- m r n ly un r h on rol o h n r l n y.

C.G. Omidyar (Ed.): MWCN 2000, LNCS 1818, pp. 47–57, 2000.

rr r s ns mul pl ss (M) s on o h mos p rv s v s r u M s h m us n w r n work. u k oll s ons r n r ns o M u o h n ur o rr r s ns ng n h x s n o h n rm n ls n ho n works. M ny M s h m s 2 3 h r w h or w hou rr r s nsng n lu ng h 02.11 s n r h v n propos o mprov h hroughpu ov r h o M . h s s h m s h n l oll s on vo n y xh ng ng n n- n on rol p k ( ) h n sh k pr or o r nsm ss on. h s n r n r nsm p k only r su ss ul x h ng . n oll s ons r solv y k ng o n r s h ul ng r nsm ss ons. h s pro o ols how v r nno m h l y r qu r m n s o r l-m r . h r l- m p k s m y no r nsm or h r l n us h on rol p k s s ll su r rom oll s ons n h r r r nsm ss ons r r n omly s h ul .

n ly s v r l nov l s r u M s h m s h v n propos m prov ng h o gu r n s or r l- m r suppor . h h s h m propos n r l- m no s on n or ss o h nn l w h puls s o n rgy h ur ons o wh h r un on o h l y n urr y h no s un l h h nn l om s l . h r sul ng s h m gu r n s pr or y o r l-m r n prov s roun -ro n s rv n oun ss l y o r l-m no s. no h r s h m 7 us s w n ow spl ng pro o ols w h l m k s ns ng o o r suppor or l n s o r l- m . ow v r h wo s h m s m n on ov o no ons r h x s n o h n rm n ls. h wo o h r r s rv on- s s h m s 9 n gu r n h on n on-r r nsm ss on s on h x h ng o h nn l s n orm ons mong h v no s. O ours s h w r l ss sp rum s pr m um h r qu n x h ng o n work n orm ons wh h m y us mu h ov rh shoul vo .

n h s p p r w n ro u ok n ss ng r ( ) s ru ur or n oor ho n works n wh h rm n ls h v low mo l y us o low mov m n sp (. . w lk ng sp) n l m mov m n sp . h poss l s n ros n lu m ng rooms w r l ss o s n o h r n oor nv ronm n s wh r w r ng s o k p o m n mum. mpor ry ho n work n r p ly ons ru n us y us rs o x h ng n orm ons n su h s u on. h propos M s h m s s on m ok n pro o ol 10 . h m ok n pro o ol s su l or r l- m ommun ons no only us o s onn on r n ur u lso u o h h h s h mpor n prop r y o oun ss m wh h s n ss ry or r l- m ommun ons. s us p rv s v ly n h w r n work. ow v r h h ll ng s on us ng h ok np ss ng s ru ur n h w r l ss ho n work om s rom hr sp s h unr l l h r r s s o w r l ss m um h h n rm n ls n h yn m s o no s. h yn m s o no s m ns h h no m y sw h o or on unpr ly. r ss ng h s ok n ss ng r s ru ur s op wh h s mor s ly m n n h n r on l ok n p ss ng s h m . ur h rmor m h n sm s prov o r ons ru h wh n h ok n p ss ng pro ss s h mp r u o h los o h ok n. s long s no m n ns s pos on n h o r qu r m n s o h s no n gu r n .

n on 2 w s r h M s h m n r v h on on o gu r nh oun ss l y o r l- m no . h m n n n o s lso r ss n h s s on. h n on 3 pr s n s h p r orm n v lu on o propos s h m . n lly w on lu h p p r n h l s s on.

2 MAC Description

h w r l ss ho n works un r ons r on ons s o no s pprox m ly l ss h n 30 ll o wh h r un on lly n l. h no s ry ng o ommun w h on no h r ov r s ngl hop or w h ss po n or lus r h o r h h n no s. n h s s h s n r s w h n s gh o h ss po n or lus r h . h s s r son l ssump on or h h r r h l ho n works n wh h h n work no s r p r on n o lus rs. s p n gur 1 yp l s ru ur looks l k - r on l r w h h ok n-ro on g nn r ( ) ng h roo . h r r wo log l ok n p ss ng r ons orw r n kw r . h orm r s us y p r n -no orw r ng ok n o s h l -no s wh l h l r s us y h l -no r urn ng h ok n o s p r n -no . h m n n n o s pr s n n on 2.2. om p r m rs us n h ollow ng s uss ons r n low.

- s h prop g on l y w n p r o no s wh h s suppos o n l o ll no s.
- s h r nsm ss on m o ok n r m .
- s h o l r s rv n w h o ll no s n . s h n w h r s rv y no .
- s h ommon oun ss l y o no s n . s h oun ss l y o no .
- s h o l num r o no s n .

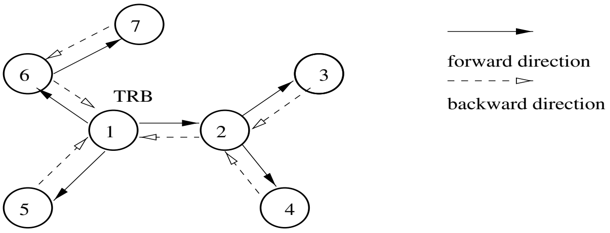


Fig. 1. yp l s ru ur o

ok n r m on ns h ollow ng n orm on ok n s n r ( ) wh h r pr s n s h n y o h no r nsm ng h p k ok n r v r ( ) wh h r pr s n s h n y o h no nv o r nsm n x n om ss gm n ( ) fl g wh h s us n h pro ur o no jo n ng h on gur on p r m rs o ( ) wh h s us n on gur on. h ok n r m n rr n h p ks o r s h ov rh us y h ok n r nsm ss on. h n urr n ok n-hol r r nsm s l s p k s h o h no nv o r nsm n x n h s p k . h ok n-hol r h s no p k o s n n v u l ok n r m s r nsm o h n n r v r n . h o h no s g v n wh n jo ns h . r no r v s h ok n rom s p r n -no rs s wh h r h fl g s s . h fl g s s h s no mus w r n om ss p r o wh h s us y n w no s o su m r qu s o jo n . h n g ns o r nsm s p k . h fl g s yn m lly s o h v h h gh h nn l u l z on. n h s s h m w n h h fl g s s on v ry wo y l s. n h no s n k h r urns o s h fl g n h ok n r m .

2.1 Bounded Access Delay with TPT

s on h m ok n M pro o ol orm n prov oun ss l y wh h s n ss ry o nsur h h r nsm ss on l n o h r l- m r s s s . h op s h m h nsm us y h m ok n pro o ol o llo v l l n w h o h no s. h m ng prop r s o h m ok n pro o ol shows h h v r g ok n ro on m s oun y h rg ok n o on m ( ) n m x mum ok n ro on m nno x w h . ur ng h n work on gur on pro ss ll no s n go ommon v lu or h oun ss l y n mpor n p r m r wh h gu r n s h r nsm ss on l n o h r l- m r . n h no h s r n oun ss l y r qu r m n s (no s or no) o s s h n go oun ss l y shoul h m n mum v lu o ll h . h s

$$- \quad - - \Psi \quad - \tag{1}$$

wh r Ψ s h s o h no s n .

h no n r s rv n moun o m (no s or no) or r nsm ss on v ry m r v s orw r ok n. o gu r n oun ss l y o ll no s s on h m ok n pro o ol h sum o llo o h no n no x h rg ok n o on m wh h qu ls h h l o . h r or h n w h llo on s h m n xpr ss s ollows

$$\sum_{\Psi_{TPL}} \quad + 2(\quad - 1)(\quad + \quad) + \quad - \frac{}{2} \tag{2}$$

On h no r v s h orw r ok n h m n m l m n us s . h n n us s h l n w h only h m l ps s n h pr v ous ok n rr v l s l ss h n h v lu $\frac{}{2}$. h moun o m (no s) ok n hol r n us s

$$\left\{ \begin{array}{l} +\frac{}{2}-(\quad - \quad)\ (\quad - \quad)-\frac{}{2} \\ \qquad\qquad\qquad\qquad\qquad \text{ls} \end{array}\right. \qquad (3)$$

wh r s h orw r ok n rr v l m o no s h orw r ok n l s rr v l m o h s m no .

2.2 TPT Configuration

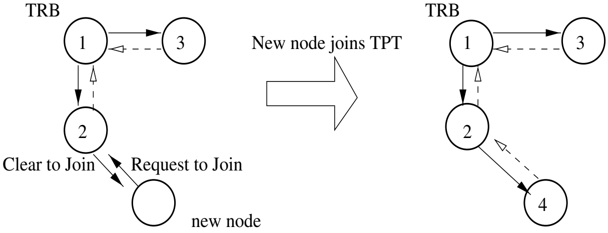


Fig. 2. o jo n ng pro ur

Call Admission Control h n no om s v rs l s ns o h h nn l. h r s no un rgo ng h s no n s n ons ru w h on no . O h rw s jo ns h urr n s n w no . h n w no n r qu s o jo n h only ur ng h r n om ss p r o ( ). h ur on o r n om ss p r o s long r h n h sum o h l ng h o r qu s o jo n ( ) n l r o jo n ( ) plus m x m l roun - r p prop g on m . gur 2 shows h pro ur o n w no jo n ng . uppos now h no 2 n r urns h ok n o s p r n -no 1. h no 1 r v s h ok n n w s or n w no jo n ng . lmos s mul n ously h n w no n p ur h s ok n us h r nsm ss on o ok n s ov r r o h nn l n g h r ss o no 2 rom h ok n. h r r s n s o h no 2 ur ng h . h rr s h n w h r qu r m n o h n w no ( ). h no 2 (l n ng no) r v s h rom h n w no n p r orms h ll m ss on on rol. h s on rol rm n s wh h r h n w ll s m or ng o h r qu r m n o n w no su j o h oun ss l y gu r n y h urr n . n w m ss on n o ur n only h oun ss l y o h urr n n m n n . .

$$\overbrace{\qquad + \qquad}^{new} +2(\overbrace{\quad +1}^{new}-1)(\qquad + \qquad)-\frac{\overbrace{\qquad - \qquad -}^{new}}{2} \qquad (\ )$$

v n h h on on s s s h no 2 s n s l r o jo n ( ) o h n w no n un qu s n ow w h h n w no . h n h no 2 nvok s h on gur on pro ss ( ) o n orm h r m n ng no n wh n r v s h ok n rom h no 1 h n x m . h ok n s s o on gur on ok n wh h rr s h n w n . h fl g n h ok n s s l wh h pr v n s o h r n w no s rom s n ng h ur ng h . h go s long w h h norm l ok n p ss ng p h. ll o h r no s n n g h n rom h on gur on ok n. r h ok n n sh s s ro on h no 2 orw r s h ok n o h n w no n h n w no s s n h n w . h pro ur o no jo n ng s p r ll l w h h norm l ok n p ss ng op r on wh h h s no mp on h p r orm n o h s s h m x p h ov rh n urr y h . s poss l h wo or mor n w no s s n h r qu s o jo n h s mul n ously. hus h s s n y h s no s w ll oll . o r solv h oll s ons h s r r n omly s h ul o r r nsm h n w no s o no r v h s rom h l n ng no n .

2.3 Token Loss

ommun on lur s n h w r l ss nv ronm n r r l v ly r qu n o urr n . h h propos M s h m h ommun on lur s m y us y h ok n p ss ng lur s u o r nsm ss on rror n or no ’s pow r ng o . or h ok n p ss ng s ru ur o s l n su h nv ronm n s h L shoul r ov r w hou us ng mu h ov rh . n h s o x n work oor n or h poll ng s h m s op . h l n s r u sys m h no n s o h v ommon s on on who shoul r spons l o r ov r h L n h s o ok n p ss ng lur . or h s r son h s s h m op s m r o h ok n p ss ng lur s n nsur h h L s r ov r . h s m r s s y h no n s r u m nn r. h llo on s h m g v n n on 2.1 n gu r n h n rv l w n wo su ss v ok n’s rr v ls no shor r h n h ommon oun ss . hus no s s m r qu l o h m r v s h ok n. h n oun s own un l h ok n rr v s h s no n x m . h ok n rr v s h s no or s m r xp r s s r n l z o n n l g n(s r ng h oun ng own pro ss). h ok n o s no rr v or s m r xp r s (. . 0) m ns h h ok n p ss ng lur h pp ns. h s no r g n r s ok n n on nu s h ok n ok n p ss ng loop. r v s h ok n n h ollow ng loop h s r ov r . O h rw s h l nk w n h s no n s p r n no s ons r ng rok n. h n om s n ro s s h s m ss g o h o h r no s. h ro s m ss g n r h ll h n no s o h s no w h h h lp o lus r or h ss po n . h n h or g n l s h m s sw h o h on n on s M s h m h no n jo n h s n w no us ng h x h ng o w h h no s n n w . us h no h s s un qu m r only h no wh h rs s h ok n los s hos n s .

ons r ng h s o n work w hou h lus r h ro s m ss g m y no r h ll no s n or h r m r xp r . hus s poss l h wo or mor no s om h us h r m r xp r . h n s v r l s m y ons ru y h s s. ll h s ssu mul - . h mul - lso x s n h mul -hop ho n works. h h mul - h r sp r us n h v . u h s m x s h pro l m h h ok n ro on o on m y n r r w h h ok n ro on o s n gh or's . h r or h m n n n o h mul - s mu h mor ompl x op n pro l m n s my un rgo ng work.

3 Performance Evaluation

h h nn l p y o h nn l n w r l ss -ho n work s l ss h n 10M s wh l h m ok n pro o ol s g n r lly op n h h gh n w h w r n work(100M s). hus w h v s mul h s s h m ov r v r y o n work on gur ons. h ov r ll p r orm n o h s s h m n h s y s w h m x popul on o n r l- m no s s s u w h s mul ons. m nly nv s g h o h wo p r m rs h oun ss l y o r l m r n h p k s z . h r r 10 r l- m no s 20 no s n our s mul on mo l. l 1 shows h r mo l o r l- m n no s. h h nn l r o sys m s 2M s. h r rr v l r o h r l- m no s k s. ssum h r l- m p k s r pr s n o h M l y r p r o lly. hus h ss m l p k s z ( ) s h pro u o h n r rr v l r ( ) o h r l- m p k n h sour r . h oun ss l y qu ls h l n o r l- m p k wh h qu ls . h r s rv n w h o h r l- m no qu ls h r nsm ss on m o r l- m p k . n l ul h h o l r s rv n w h o r l- m no s s l ss h n $\frac{}{2}$. h o h r p r m rs r shown s ollows. h s z o ok n r m s 20 y s. h s 0.1ms. h prop g on l y s 0.1 s. h lo s n s wh r s h num r o s on. us h p k s z s x h l y s n s h ss l y. h hroughpu o r s m sur s h norm l z p k s r nsm ov r h h nn l.

us h r l- m no s r gu r n w h oun ss l y n h s y s s shown n pr v ous s on shows h h m x m l r l- m p k l y s l ss h n h l n o p k . o r l- m p k s r ropp v n n h gh lo . gur 3 shows h v r g r l- m p k l y g ns h lo . h ss m l r l- m p k s z s 1 00 3200 n 00 r sp v ly. or ngly h oun ss l y s 2 ms 0ms n 100ms. h r sul shows h h v r g r l- m p k l y s l ss h n on h r o h l n o h p k v n n h h gh lo . v n h s m lo h l y o r l- m p k n r s s s s z n r s s. gur shows h hroughpu o r g ns h lo . h on n on r

Parameters	Real-time node	Data node
Interarrvial pdf	Constant	Poisson
Interarrvial rates(ms)	A_r	λ
Packet length pdf	Constant	Constant
Mean packet size(bytes)	b_{rpkt}	b_{pkt}
Reserved bandwidth	b_{rpkt}/r_c	0
Packet deadline(ms)	D	no requirement

Table 1. r mo ls

n ur o ok n p ss ng s h m nsur s h v ry r nsm ss on s su ss ul n s or on r h nn l. h hroughpu n h r or r h m x m l v lu n s no gr un r h vy lo s. n s rom gur 3 n h h r s r o w n h o r qu r m n s o r l- m r n h hroughpu o h r . h hroughpu o r r s s s h oun ss l y r s s wh n h n work s ov r lo wh l h sm ll r v lu o h oun ss l y m ns r o h h r l- m r n g . h r son s h h r nsm ss on o p k s ons r n y h oun ss l y 11 . m l rly gur shows h mp o gu r n ng pr or y o r l- m r on h v r g l y o p k . h v r g p k l y n r s s mor qu kly w h h l rg r s h lo n r s s.

gur n 7 omp r s h v r g p k l y n hroughpu o r w n h s s o r n x -l ng h p k . o s rv h h p r orm n o h s s h m s r wh n h p k r shor r. h shor r p k s mor s r l o h v r p r orm n .

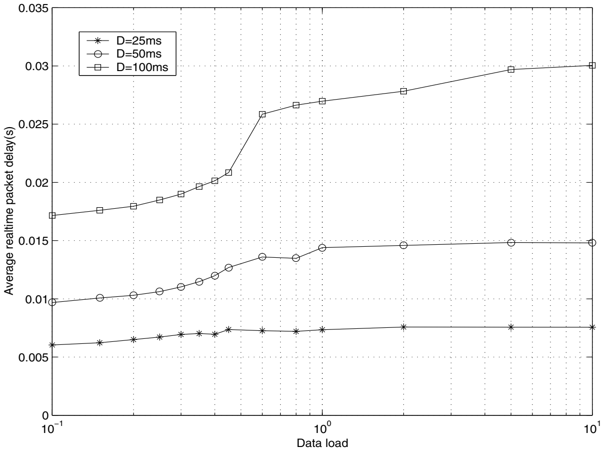


Fig. 3. l- m p k l y v rsus lo

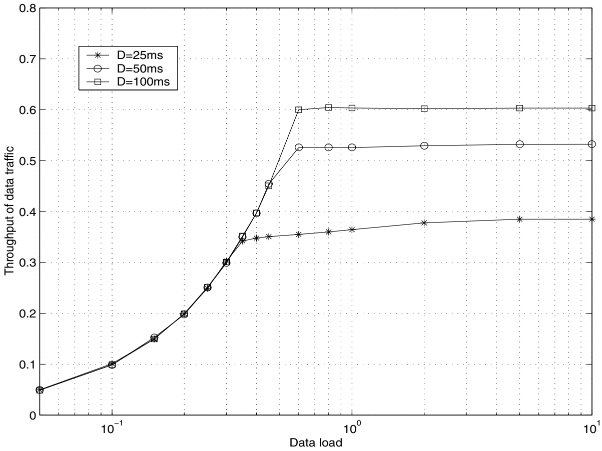


Fig. 4. hroughpu o r v rsus lo

4 Conclusions

h v s r h s ru ur o h ok n p ss ng s M s h m or ho n works n nv s g s m jor p r orm n hrough s mul on. h r sul h v shown h h s s h m n v ly suppor r l- m r n h s y s . h ongo ng work s o nv s g h mp o h ok n los v n s on h s s h m n o h l p r orm n n lys s o h s mo l. Our u ur s u s w ll o us on mprov ng h p v y o urr n M s h m .

References

1. M. Conti, C. Demaria and L. Donatiello.: Design and Performance Evaluation of A MAC Protocol for Wireless Local Area Networks. *ACM Mobile Networks and Application,* Vol. 2, No. 1, 1997, pp69-87.
2. P.Karn, MACA-A new channel access method for packet radio.: *Proc. ARRL/CRRL Amateur Radio Ninth Computer Networking Conf,* 1990.
3. F. Talucci and M. Gerla.: MACA-BI (MACA By Invitation) A Wireless MAC Protocol for High Speed Ad hoc Networking. *The 6th IEEE International Conference on Universal Personal Communications,* 1997, pp913-917.
4. C.L. Fullmer and J.J. Garcia-Luna-Aceves.: Floor Acquisition Multiple Access (FRMA) for Packet-Radio Networks. *Proceedings of ACM SIGCOMM,* 1994.
5. IEEE 802.11. Draft Standard for Wileless LAN. P802.11D5.0, 1996
6. J.L.Sobrinho and A.S.Krishnakumar.: Real-Time Traffic over the IEEE 802.11 Medium Access Control Layer. *Bell Labs Technical Journal,* Vol: 1, No.2, Autumn 1996, pp172-187.

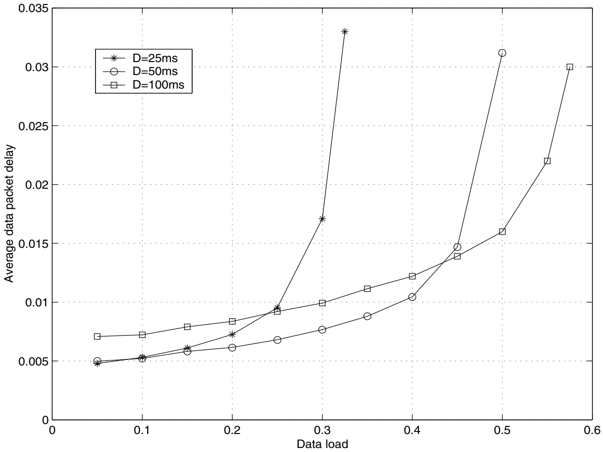


Fig. 5. p k l y v rsus lo

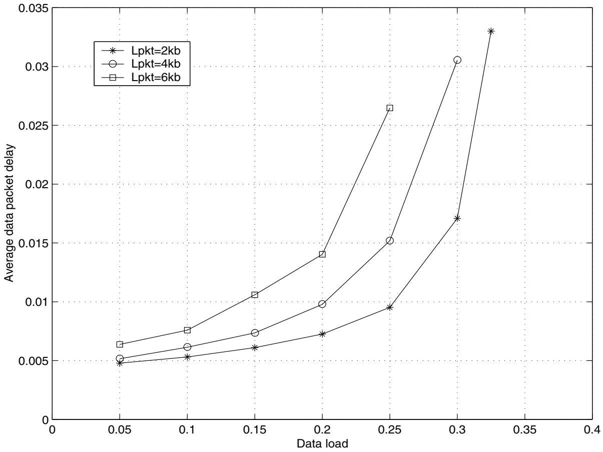


Fig. 6. p k l y v rsus p k l ng h

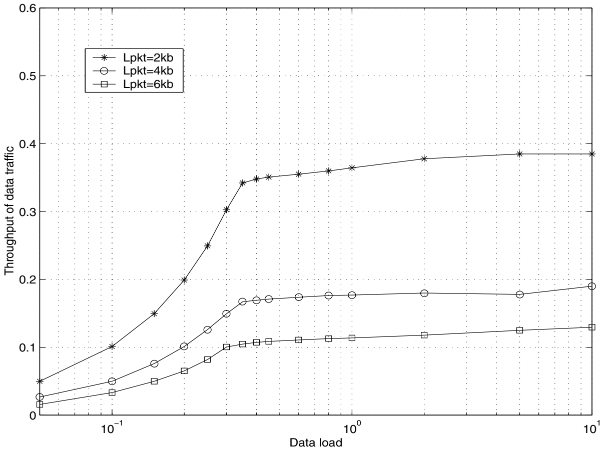


Fig. 7. hroughpu o r v rsus h lo

7. Michael J.Markowsk and Adarshpal S.Sethi.: Fully Distributed Wireless Transmission of Heterogeneous Real-Time Data. *Vehicular Technology Conference,* Volume: 2, 1998, pp1439-1442.

8. C.R. Lin and M. Gerla.: Asynchronous Multimedia Multimedia Radio Network. *IEEE INFOCOM'97,* 7-11 March, 1997.

9. A.Muir and J.Garcia-Luna-Aceves.: Supporting real-time multimedia traffic in a wireless LAN. *Proc. SPIE multimedia Computing and Networking,* pp.41-45 1997.

10. R.M.Grow.: A timed token protocol for local area networks. *Proc.Electro/82,Token Access Protocols,* May 1982.

11. Seung Ho Hong.: Approximate analysis of timer-controlled priority sceme in the single-service token-passing system. *Networking,IEEE/ACM Transaction,* Volume:22, April 1994, pp.206-215.

Challenges for Mobile Voice-over-IP

t im w l[1] y - n n[1] n o m . n n[2]

[1] Applied Research, Telcordia Technologies, Morristown, NJ 07960, USA
{pagrawal,jcchen}@research.telcordia.com
[2] Department of Computer Science, University College Cork, Cork, Ireland
cjs@cs.ucc.ie

Abstract. The use of IP to transport voice is receiving significant attention in the telecommunications and data networking industries. Commonly known as Voice-over-IP (VoIP), this technology has the potential to allow user communication involving multiple media types, and between terminals that offer improved capabilities and user interfaces. Meanwhile there has been explosive growth in the demand for wireless connectivity, raising the question of how VoIP can be extended to situations where users and terminals can be mobile. In this paper we survey the technologies necessary to provide mobile VoIP, and identify relevant technical challenges.

1 Introduction

t l p ony i xp t to ontinu owt in uppo tin o po t n omm i l vi . i owt point to l u in n m k t v lu o oi -ov - (o) in t n utu . u ntly t l p ony p ovi voi vi to n t min l t t tt to wi n two k . it t p oli tion o mo il n wi l vi n mo u i onn t om t i fix point n om mo il t i n in in n to pt o to t mo il n wi l om in 10 .

ow v t mo il nvi onm nt i mu mo yn mi n u j t to n t n t t ition l wi lin nvi onm nt. un t inti o t wi -l n mo il nvi onm nt ll o n in l v l o pt ility. o mo o u t l-tim ommuni tion i n lin p oto ol i vit l in p ovi in i ly li l n o u t onn tivity in u ommuni tion nvi onm nt. n i-tion to t li in n l in ll i n lin p oto ol m y l o n to monito n m int in onn tivity w n t n t min l i movin n /o t t n mi ion p iliti v yin . lin wit i u o u n t min l mo ility in n int n two k t k on p i l i nifi n w n voi t n po t i involv u o l t n y n n l p o m n on t int t t not impo t nt o non- l tim ommuni tion. imil ly t vi qu l-ity n y o voi ommuni tion m n p k t ulin n ou v tion t niqu w i n op t tiv ly in t p n o mo ility. i p p i o niz ollow . tion 2 p ovi u v y o i n lin o o n i i n lin it tu o u wit mo il o t min l .

C.G. Omidyar (Ed.): MWCN 2000, LNCS 1818, pp. 58–69, 2000.

tion 3 xpl in ow u n t min l mo ility n l in t t l p on y t m n in t nt n t. o m n i u o uppo tin voi to mo il t min l i u . tion 4 i u ow to p ovi uppo t o qu lity o vi (o) o p k t voi ommuni tion ov wi l link n w t min l non- t tion y. in lly tion 5 on lu t p p .

2 Signaling

fi t ll n o t l p ony i to initi t int tiv ommuni tion ion tw n u . wo m jo t n v ntly m o t i n lin n ont ol o o . fi t i .323 8 uppo t y t - n t on i t ion niti tion oto ol ( ) p opo y t . two p oto ol p ovi imil i n lin un tion lity ow v .323 i wi ly ploy in omm i l p o u t . i o o t l o x mpl N t tin .323- ompli nt p o u t in 199 . n t ot n i li tw i t p oto ol w i p ovi impl impl m nt tion n t o i n y t n .323. two p oto ol ow v u ntly o not p ovi uppo t o ou v tion n o u nt . (i t i ut p n i n lin it tu) 5 n (l p ony v k t n two k) 9 p opo y p ovi m wo k to int t ll i n lin n ou m n m nt o t l p ony n to n l tiv n w vi o p k t t l p ony. l o i - o i joint p oj t im on p ovi in i n lin lon wit t min l mo ility n o uppo t.

2.1 H.323

i in lly .323 w v lop o vi u l t l p on t min l on n in ov non u nt o L N n i n um ll t n ov in u io n vi o o ll i n lin onn tion ont ol t n on n ont ol m i t n po t t . n .323 t i n lin un tion lity i mi t to n t min l w i int lli nt n point in t o t um n point u in t N (u li wit l p on N two k). .323 i l o not ti to in l t n po t m ni m n n un ov v i ty o n two k in lu in n N.

i . 1 ow typi l it tu n ompon nt o n .323 L N w i i ll .323 zon . t min l in .323 u u lly i multim i . t w y ( ) i n n point on t n two k w i p ovi o l-tim two-w y ommuni tion tw n .323 t min l on t p k t- n two k ot t min l on wit i uit n two k n ot .323 t w y . ultipoint ont ol nit ( ) i n n point w i p ovi t p ility o t o mo t min l n t w y to p ti ip t in multipoint on n . t k p (K) i n ntity t t p ovi t n l tion n ont ol to t n two k o .323 t min l t w y n . t k p m y l o p ovi ot vi to t t min l t w y n u n wi t m n m nt n lo tin ot t w y .

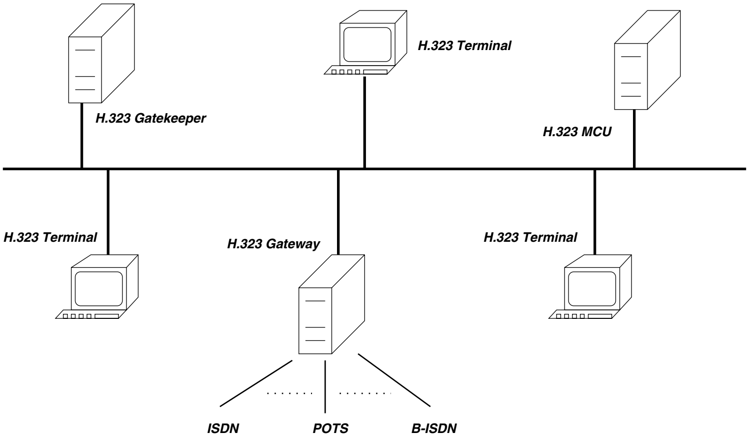


Fig. 1. .323 zon

.323 u .225.0 t onn tion t li m nt p oto ol n .245
t ont ol p oto ol tw n .323 li nt to t li ll n oti t t min l
p ility n op n lo i l nn l . n t ll t t involv t k p
t min l t t w nt to t up ll fi t n qu t to it K. t t K
x n m wit t ot K t K pon to t ll o i in to
(ll) wit t o t i tin tion (ll). ll t n n
t qu t i tly to t ll . ll t n k it K o p mi ion t
n in ll p o in m to t ll . n t p mi ion i nt
y it K t ll pon to t ll . t t n oti tion tw n t
n point t m i nn l op n n t t min l t n x n
m i t u in / .

o p ovi t min l mo ility o .323- t l p ony 11 p opo
l v in t yn mi join n p tu o multipoint on n wit lo tion
up t . t uppo t mo ility o .323 t min l wit out in n w ntiti
n wit minim l mo ifi tion to t t n .

2.2 SIP

(ion niti tion oto ol) i i n lin p oto ol o t l p ony v l-
op y t . o i in t out o li tw i t nt n t pp o n
u m ny o t fi l n o in ul o o n ut nti tion
m ni m o . wit op t in p n ntly o t p k t
l y n u t xt to n o it m w i ilit t p in n lp
to impli y u in .

o pl ll in t ll o i in to (ll) lo t n pp op i t
v typi lly y N . ll t n n n N qu t to t
v . n lin o N p n on t n tu o t v . n t
o p oxy v t v qu i lo tion v to fin t o t

i tin tion (ll) n o w t N m to t ll . ll n n K to t p oxy v w i o w it k to t ll . ll u qu nt i n lin tw n ll n ll i t n n l t ou t p oxy v .

n t o i t v t v qu i t lo tion v to fin t ll t n uppli t lo tion o t ll to t ll . ll t n i u n N m to t ll . ll u qu nt i n lin i i t. n l o u imply om li nt to not li nt. l wit u mo ility y lyin on lo tion v ut w not i n to t i u o t min l mo ility. tion 3 i u olution t t u to p ovi t min l mo ility.

2.3 DOSA and TOPS

ition l i uit t l p ony vi n u t t n -to- n ou v o t ll p ty p on i m to in . ow v ot .323 n i n lin p oto ol wit out ny on i tion o ou v tion n o u nt . (i t i ut p n i n lin it tu) w i i m wo k o ll i n lin n ou m n m nt o t l p ony p o-po y p ovi imil vi to onv ntion l t l p ony. n t l p ony li nt u o multim i t min l pto p ti ip t in n -to- n p ility n oti tion ll i n lin n ou v tion. in o po t n xpli it oo in tion tw n t ll i n lin p oto ol n t ou m n m nt p oto ol u t t u ut nti t n ut o iz o ivin to t n n o o i t wit t t l p ony vi .

(l p ony v k t n two k) l o p opo y on t ot n llow u to u mo il t min l o to mov tw n t min l ut till n y t m u i tin ui in n m . n ont ol ow ll out to t m. min l in v n o p iliti to uppo t i nt m i . m jo l m nt o in lu i to y vi n ppli tion l y i n lin (L) p oto ol lo i l nn l (L) t tion n m ni m to uppo t v i ty o on n in mo . i to y vi p ovi fl xi l m ppin o u n m to t o t min l w t u n . L n oti t p iliti t li n m int in ll t t n uppo t v n tu . L t tion i ol t t ppli tion om t un lyin n two k . y t m n impl m nt n i in u xp im nt lly wit in L - .

2.4 ITSUMO Signaling

(nt n t nolo i uppo tin niv l o il p tion) l o i n o i joint p oj t nvi ion n n -to- n wi l / wi lin pl t o m o uppo tin utu l-tim n non- l-tim multim i vi . o l i to u n t i (n ov) n tion wi l p k t t nolo i to i n wi l pl t o m t t llow mo il u to

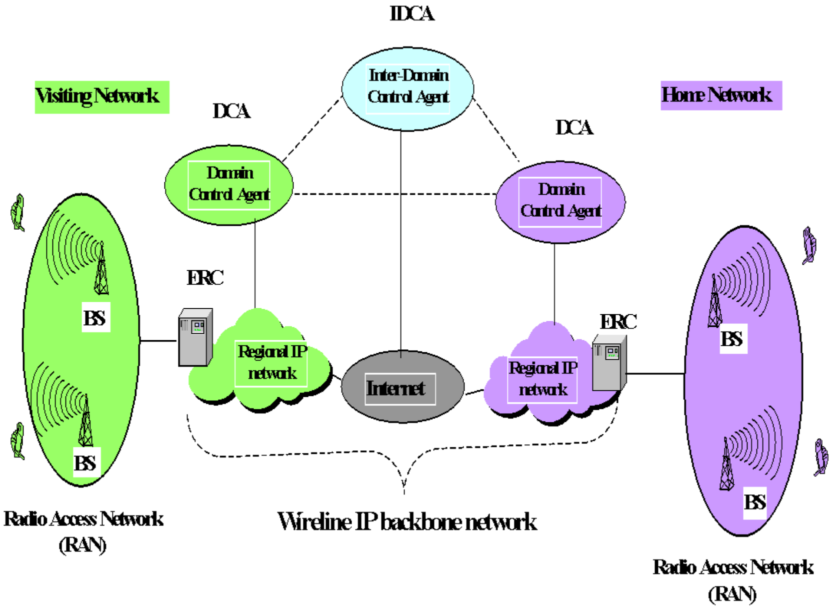


Fig. 2. lon t m n two k it tu

ll vi on n xt n tion nt n t. i . 2 pi t t n -to- n p k t pl t o m o ll wi l /wi lin n two k w i omp i ll wi l n two k n n k on n two k. k on n two k i n n -to- n wi lin in t u tu t t will omp i ion l p ovi wi lin n two k t t onn t t ou t nt n t. i mo il t tion /t min l wi l n two k l o omp i io n two k (N) n n out n ont oll ( ). n o to uppo t t n o it u wi l n two k int t wit t n two k ont ol ntiti t t own om in ont ol nt ( ) in i . 2.

lon -t m ont ol it tu o i own in i . 3. t u o it n -to- n ont ol (. . t .) it tu . ll mo il t tion ( ) n fix o t v u nt t t int t wit t v (i. . p oxy v i t v n i t) o t / wit in t n two k. n i . 3 t v ntity o ion l n two k p nt t t o p oxy n i t v wit in t ion l n two k t t p o m t n two k ont ol n i n lin un tion . imil ly t i t p nt t v n t (o t o v n t) t t pt (pt) qu t n p ovi (p ovi) lo tion vi t t imil to t o o t om /vi itin lo tion i t i (L / L) in to y llul t l p ony. i . 3 ow t (mo ility ut nti tion ut o iz tion ountin n o) ntity i uilt on top o

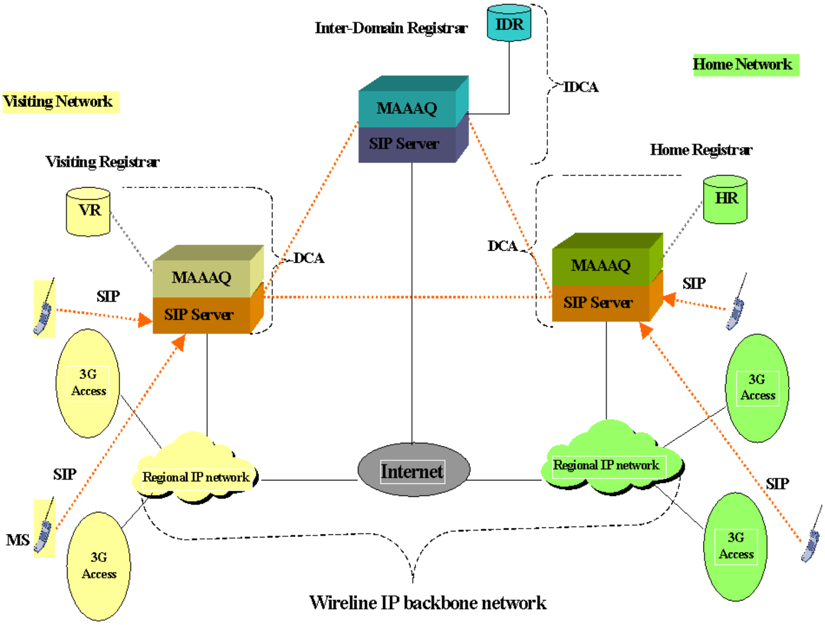


Fig. 3. lon t m i n lin it tu

- (n /o) y t m n u t lo tion n i n lin vi o to uppo t o min u .

3 Mobility

i u o mo ility n i u y fi t on i in it impl m nt tion in t t l p on y t m n t n u n t min l mo ility in t nt n t.

3.1 Mobility in the Telephone System

n t l p ony y t m un m nt l qui m nt i to l to lo t u in o to t li l-tim ommuni tion. onv ntion l t l p ony o not p ovi m n o lo tin u p - ut t um l tiv ly t ti o i tion o fix t min l wit u n li on t ll to on ult i to y to m p u n m to t min l num ( ). u in ll n t mpo ily i umv nt t i m y n lin t ll o w in tu llowin ll ll to iv n t min l to i t to n lt n t t min l. n mo t t l p on y t m it i l o po i l to t o- ll p on l num - t l p on num o i t wit n in ivi u l u t t n iv n t min l. y t m m int in yn mi m ppin tw n t p on l t l p on num n t num o p ifi t l p on . i m ppin i up t xpli itly y t u to w om t num i i n .

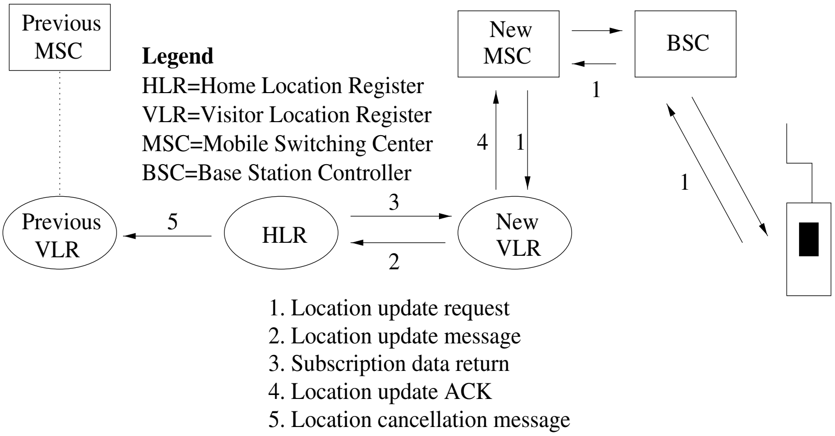

Previous MSC
Legend
HLR=Home Location Register
VLR=Visitor Location Register
MSC=Mobile Switching Center
BSC=Base Station Controller
New MSC
BSC
1
4
1
1
Previous VLR
5
HLR
3
2
New VLR
1. Location update request
2. Location update message
3. Subscription data return
4. Location update ACK
5. Location cancellation message

Fig. 4. o ility in t llul l p on y t m.

llul t l p ony u t m i mo l onv ntion l t l p ony xpt o ou t t t min l mo il n llow to o m tw n ov (ll). o t li ll t y t m mu t fi t lo t t u nt lotion o t t t t min l. i i i v y in t t t om lo tion o t t min l t min y t t min l num . i t u u lly ll t om Lo tion i t (L) m int in n nt y i nti yin t t min l l t o lo tion. ov it own t t t m int in nt i o ll vi itin t min l - u u lly known t i ito Lo tion i t (L). i . 4 ow t t p involv in t i t min l lo tion p o . t min l mov om on ov to not in p o known n o n nt y i t in t n w L t nt y in t p viou L i mov n t L i up t to point t t n w ov . oul t t min l involv in ll w n t mov tw n ov o u t n o m ni m n o t voi onn tion to t n o to t n w ll.

3.2 User Mobility in the Internet

n t nt n t t o i tion tw n u n t min l num i not um to t ti . xp t to l to t nt n t om ny omput n it i ommonly t t t t t min l num ( ) i n to iv n t min l i n t yn mi lly u in t yn mi o t onfi u tion oto ol ( ). u to lo t t o t u u nt t min l qui t t yn mi in in m int in n on ult w n v on wi to lo t t t u . ly o p o u t li on t u o p i l n zvou vi u n nt n t l y t ( ) v to llow u to ow t li t o u onlin n o t u nt o i t wit u . tim t o li nt o tw x ut it took o t i t tion p o u on l o t u .

o nt pp o in lu t .323 t k p w i p ovi m ppin tw n u li u n m il- n t o pon in t n po t . o n m il to i nti y u p ovi n min p w i i onv ni nt l l n n p ovi uniqu i ntifi tion. .323 li nt o tw n t k o i t in t u wit t k p . Lo tion v p ovi t m un tion ut in ition to xpli it i t tion t v n l o onfi u to u ot t niqu to t y to lo t u in lu in o x mpl t u o t fin utility to p o t o lik ly omput . i to y vi llow u i t tion ut in ition uppo t t o i tion o mo t n on t min l wit iv n u n ll p o in lo i to l t mon t i nt t min l on tt i ut u ll i ntifi tion tim -o - y n m i typ i .

3.3 Internet Terminal Mobility

o t mo ility to t itu tion w t min l n mov om on lo tion to not . two to on i . Nom i omputin i w n t min l i u in m ny i nt lo tion ut in o t in lo l vi . li nt point i t t ny nt i in u i to y v mu t up t wit t n w . ot mo l i mo il omputin w t min l i llow to t in it no m tt in w i lo tion it pp n to op tin . i i t p o l m t t t o il p oto ol 12 olv .

o il - int o u t on pt o t min l vin two om n -o . om to t t t min l i i n in it om u n t. -o i t mpo y i n o u y t t min l w il it i vi itin not u n t o it n tim t t min l mov om on u n t to not . i i i ollow . n t min l mov out i it om u n t to not u n t it fi t qui lo l -o n i t t t wit om nt k on it om u n t. k t to t t min l ontinu to iv on t om u n t pi k up y t om nt n o w to t t min l t it u nt -o (po i ly vi o i n nt). i i own in i . 5. n t min l mov tw n u n t t i n o p o u t t t k o up tin i t tion in o m tion t t om nt.

v l p o m n i u i w n on i in t u o o il - o uppo tin o . i t t i t i u o n -to- n l t n y w i o voi t l p ony mu t ti tly on t in . n t i p opo l o o il - ll p k t tin o mo il o t t v l vi t o t om u n t - ultin in itu tion known t i n ul outin . l ly t i n ult in i nifi nt in in t n -to- n l y o o ll. olution to t i n ul outin xi t out optimiz tion p oto ol w i llow o- ll o pon in o t to l n t u nt -o o t t t o t. i p oto ol i i n n optimiz tion n n ult in l t n y w i i i ly v i l .

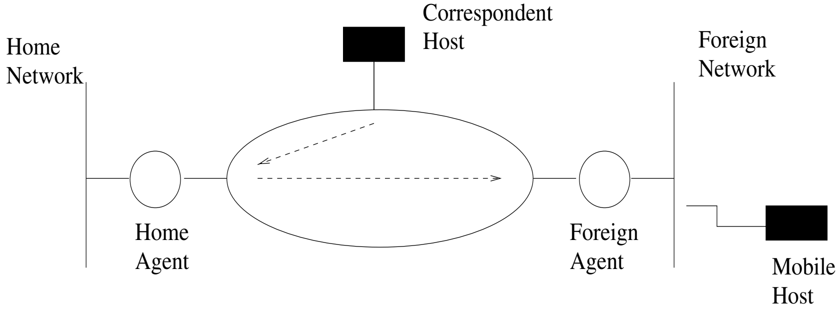


Fig. 5. o il - op tion.

not p o m n i u i t l t n y o n o in o il - . u o t n to int t wit t om nt in o to p o m n o t l t n y involv n i nifi nt. v n v l p opo l i n to imp ov t p o m n o n o u 2 . i p opo l int o u i y w i llow v l ov to m n oup wit in w i o t mov m nt i invi i l to t om nt n o pon nt o t . l t to t t i t llul wo k 3 w i u t t u o lo tion m n m nt t niqu imil to t o u in t llul t l p on y t m o p ovi in l l mo ility m n m nt olution wit o il - . in lly 13 p opo u in to p ovi low-l t n y mo ility o t min l involv in multim i ion . n t t olution i n lin m - out u u l on t op tion o o il - ut t m i t m - out t t ppli tion l y y vin n oti t t v iou t m to u t -o .

4 Service Quality

n ition to mo ility- l t p o m n i u i u ov o om-muni tion involvin mo il t min l qui two k y vi qu lity i u to t ility to ul voi n non- l tim t on wi - l m ium n p oto ol to p ovi ou v tion in itu tion w on o mo t min l n mo il .

4.1 Medium Access

i u o p ovi in ul m ium o ou iv on i - l int t in t lit tu ot o wi lin n wi l n two k . opul wi l lo l n two k v ollow ont ntion-mo l omp l to t t o n t n t n two k m kin t m n lly un-uit l o multim i t . o x mpl t 802.11 t n o wi - l L N impl m nt i n multipl ( ) it ult mo o op tion known t i t i ut oo in tion un tion ( ). u t wit t n t on qu n o t i pp o i n in ility to p i t l t n i in t y p n nt on t u nt o lo . 802.11 t n- l o p ovi n option l mo o op tion ll t point oo in tion un tion ( ) w i int o u n ont oll o ov . ont oll impl m nt ont ntion- p io in ition to t ont ntion- p io o t ult mo . n t ont ntion- p io t ont oll poll t o o t w i v in i t i to v p i t l ivin t m ul oppo tunity to t m ium. i nt olution i x mplifi y t i l nt t vi l t- o m ( ) 4 . uppo t mix o on t nt it t voi n t- o t t y mployin m ium p oto ol on tok n p in . n n ont oll in ov un n l o it m w i on qu t om t min l nt to t m ium y p in tok n to t t min l. o voi t t t min l x n m wit t ont oll to p i y it n wi t qui m nt n n t t t w i it will qui tok n llowin t ont oll to p o m mi ion ont ol.

4.2 Resource Reservation

ultipl vi l in n n two k n p ovi y t ou v - tion p oto ol ( ). on t notion o t flow llow n to i t t it pl n to t n mit n iv t l o vi t y i . i in o m tion i u to p o m mi ion ont ol n t li flow t t t o t int m i t out . n t i in o m tion i o t- t t i. . it tim out n i i i not ul ly . k y p o l m t t i i ow wo k in t p n o t min l t t mo il . n p ti - ul vin t li v tion i it o ot n n iv mov to i nt n two k t n n w v tion i qui . i p o m n impli tion l tin to t l y o t li in n w v tion. t l o i t po i ility o t min l movin into ov in w i t i ou not v il l .

n xt n ion to n p opo to l wit t i u . Known o il - 1 t i i to fin v tion in it mo ility in p n nt o mo ility p n nt. o mo ility in p n nt vi it i n - y to m k p ti l v tion i. . t o t lo tion t t mo il t min l mi t vi it u in t li tim o t flow. i t o lo tion mu t p ifi y t u i p t o mo ility p ifi tion. m i tin ui tw n two typ o v tion n tiv v tion to t u - nt lo tion o t t min l n p iv v tion to ll ot ll in t mo ility p ifi tion. p iv v tion om n tiv v tion w n t t min l nt t o pon in ov . u i to t mo ility p n nt vi m k v tion om t n only to t i u nt ll. o voi low utiliz tion o t n two k t ou o i t wit p iv v tion llow to u y ot flow i t y w un v .

5 Conclusions

u o to t n po t voi i ivin i nifi nt tt ntion in t t l ommuni tion n t n two kin in u t i . ommonly known oi -ov - (o) t i t nolo y t pot nti l to llow u ommuni tion involvin multipl m i typ n tw n t min l t t o imp ov p iliti n u int . nw il t n n xplo iv owt in t m n o wi l onn tivity i in t qu tion o ow o n xt n to itu tion w u n t min l n mo il . n t i p p w v u v y t t nolo i n y to p ovi mo il o . t i l t t mo ility iou impli tion o lo tin u in o to t li ll n m int inin onn tivity wit pp op i t o t ou out ll v n in t p n o mo ility. o ility l o n imp t on ll i n lin p i lly in l tion to vin to pt ll t i ti n m i tin tion t min l mov . v i u t k y i u n iv n i tion tow po i l olution .

i nin n two k to p ovi wi lin o lon i ot vi i in it l v y i ult t k n int o u o t o t ni l n op tion l p o l m . n mo il o y t m t ition l i u o t kin u t t n o m tw n i nt utonomou y t m n p ovi in ml n o tw n wi l in t u tu t t own n op t y i nt o niz tion . m in ompl x i u t t n to olv wit in t ov ll o l o p ovi in multi- vi n two k w voi t l p ony i ju t on o t o vi ll on liv y to t n -t min l.

References

[1] Badrinath, B., and Talukdar, A. IPv6 + MOBILE-IP + MRSVP = Internet Cellular Phone? In *Proc. of IFIP Int. Workshop on Quality of Service* (May 1997), pp. 49–52.

[2] Caceres, R., and Padmanabhan, V. N. Fast and scalable wireless handoffs in supports of mobile Internet audio. *ACM/Baltzer Mobile Networks and Applications 3*, 4 (1998), 351–363.

[3] Campbell, A., Gomez, J., Kim, S., Wan, C., and Valko, A. A cellular IP testbed demonstrator. In *Proc. of the 6th IEEE International Workshop on Mobile Multimedia Communications (MOMUC)* (1999), pp. 145–148.

[4] Goel, S., Mishra, P., Saran, H., and Sreenan, C. Design and evaluation of a platform for mobile packet telephony. In *Proc. of IEEE International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)* (July 1998), pp. 71–81.

[5] Goyal, P., Greenberg, A., Kalmanek, C. R., Marshall, W. T., Mishra, P., Nortz, D., and Ramakrishnan, K. K. Integration of call signaling and resource management for IP telephony. *IEEE Network 13*, 3 (May/June 1999), 24–32.

[6] Handley, M., Schulzrinne, H., Schooler, E., and Rosenberg, J. SIP: session initiation protocol. IETF RFC 2543, Mar. 1999.

[7] ITSUMO Group. Benchmarking of ITSUMO's all IP wireless architecture. Mobile Wireless Internet Forum (http://www.mwif.org/) <mwif2000.028.0>, Jan. 2000.

[8] ITU-T Rec. H.323. Packet-based multimedia communications systems, Oct. 1997.

[9] Kalmanek, C., Kaplan, A., Marshall, W., Mishra, P., Onufryk, P., Ramakrishnan, K., and Sreenan, C. TOPS: an architecture for telephony over packet networks. *IEEE Journal on Selected Areas in Communications 17*, 1 (Jan. 1999), 91–108.

[10] Kanter, T., Olrog, C., and Jr, G. Q. M. VoIP for wireless and mobile multimedia applications. In *Proc. of the 1999 Personal Computing and Communications Workshop* (Nov. 1999), pp. 141–144.

[11] Liao, W. Mobile Internet telephony: mobile extensions to H.323. In *Proc. of IEEE INFOCOM* (New York, NY, Mar. 1999), pp. 12–19.

[12] Perkins, C. IP mobility support. IETF RFC 2002, Oct. 1996.

[13] Wedlund, E., and Schulzrinne, H. Mobility support using SIP. In *Proc. of 2nd ACM International Workshop on Wireless Mobile Multimedia* (1999).

Smart Delivery of Multimedia Content for Wireless Applications

Theo Kanter[1], Per Lindtorp[2], Christian Olrog[3], and Gerald Q. Maguire Jr.[4]

{[1] Theo.Kanter [2] Per.Lindtorp [3] Christian.Olrog}@era.ericsson.se,
Ericsson Radio Systems AB, SE-164 80, Stockholm, Sweden.
[4] maguire@it.kth.se,
TeleInformatics, KTH, SE-164 40, Stockholm, Sweden.

Abstract. Packet-oriented access to cellular networks enables us to deliver multimedia content to mobile users. As cellular networks will continue to deliver circuit switched voice for some time to come, care must be taken to avoid interference between these delivery mechanisms, while maximizing the range of services and the number of users. Smart delivery of multimedia content involving agents running in the mobile, the base station and the content provider allows us to dynamically adapt the application and network behavior to each other in order to meet the criteria for specific applications. In particular, this paper examines the delivery of streaming media and interactive voice as Voice over IP (VoIP) to mobile users. Our conclusion is that this, in combination with the dynamic adaptive properties as introduced by the agents, enables us to transfer voice entirely IP over wireless links, thereby freeing further resources for the new applications that we refer to in this paper.

1 Introduction

Presently, the telecom and datacom industries are converging in different ways. With respect to mobile telephony with GSM, new devices are appearing on the market that integrate data with the telephony voice service in new ways. So-called Smart Phones either include the functionality of an organizer, or can connect wirelessly to an external Personal Data Assistant (PDA) and integrate the functionality of the organizer for smart dialing and messaging with the application that is running in the handset. The Wireless Application Protocol (WAP) is intended to move the point of integration of these services into the cellular access network. WAP-gateways can be used to adapt and convert Internet information, so that the mobile terminal can be used for interacting with a wider range of *network*-centric services (e.g. electronic payment, subscription to information services, unified messaging, etc.). However, WAP is neither intended nor well suited to transport multimedia content, but rather was targeted at simply extending GSM networks with data services. Only through WAP-gateways can these services be connected to services on the Internet. Therefore, WAP excludes mobile users from directly interacting with Internet content. On the other hand, simple low-bandwidth GPRS (General Packet Radio Service) is introduced in GSM-networks, which will provide direct Internet access to mobile users. GPRS enables the development of multimedia applications for the mobile

C.G. Omidyar (Ed.): MWCN 2000, LNCS 1818, pp. 70-81, 2000.

device. These applications can integrate content that resides on the Internet. EDGE (Enhanced Data-rate for GSM Evolution), the successor of GPRS, increases the bit-rate and thereby further relaxes the requirements on the mobile applications and Internet content, thus bring even more new multimedia applications to mobile devices. Mobile devices are now able to perform significant computations based on events from various input devices and/or information sources. These events provide information about the user's context and the conditions, which the link is facing. Therefore, applications in mobile devices and nodes in the network that co-operate to deliver multimedia services to users can adapt their mode of communication dynamically based on such events. In such cases, the requirements for the delivery of multimedia IP-content over a link using a wireless access networks are even further relaxed.

2 Problem Statement

The important question is therefore: how cleverly can we dynamically shape the applications and Internet content, in order to maximize the delivery of multimedia content to mobile users? Multimedia applications put wide-ranging requirements on links. However, the quality of service (QoS) requirements involved can be categorized by several parameters, which differ in importance for the different type of services (and applications). We examine these parameters below.

2.1 Latency

Latency is important for isochronous services, such as voice (e.g., interactive speech) where delays up to 250 milliseconds are perceived as acceptable. Beyond 500 milliseconds the behavior of users gradually adapts itself to the increase in delay and the receiving party usually waits for a ready signal before starting to send. Latency may be due to network-related delays, and buffering in either the application or the device. Latency is not critical in streaming applications, such as Internet radio and other applications that playback multimedia content (e.g. MP3-files) as long as the user is assured that the content is going to be received within a bounded maximum delay.

2.2 Robustness

Latency is also important for interactive network games, such as Quake and Unreal, but in this case delay is not correlated to congestion problems. On the other hand, it is important that the packets arrive, otherwise synchronization problems will occur (these are not critical as continuous updates of players' locations are sent). This is normally handled by TCP or by UDP and an application-specific handshake protocol. However, the link level may also provide this service. The interleaving that GPRS does increases the probability that a packet is *not* lost, but on the other hand the user pays a penalty in increased latency, even if the packet is **not** lost.

Ericsson's GPRS Application Alliance has tested the currently most popular network game, Unreal Tournament, with a GPRS simulator with good results [16]. Using only one timeslot, the latency (about 300 - 800 msec) at 26 kbps incurred by GPRS interleaving does somewhat adversely affect players' performance and appreciation of this application, but not in a critical way.

2.3 Speech Quality

The coding/encoding algorithms (codecs) dictate the upper bound of the perceived QoS of multimedia streams (such as speech, video or audio). The lower bound is dictated by the percentage of link packet-loss that the codec is able to tolerate before its performance suffers severely. Modern codecs (e.g., Voxware RT-24) can tolerate up to 30-40% packet loss, with no additional latency.

2.4 Requirements

In order to assure that a certain application is feasible we must show that we can meet the requirements in a satisfactory manner at all times and in a scalable way. Two services of particular interest that will be studied further in this paper are:

1. delivery of streamed audio (e.g., MP3-files, Internet radio stations) and
2. interactive voice - e.g. Internet Telephony

The question is how we should use the functionality in: the mobile device, the wireless link, and the network, in order to provide scalable services and applications, such that the number of users is maximized. When designing our applications, we can use the knowledge that applications have of specific end-user requirements to our advantage, specifically by using strategies that dynamically shape the applications and their use of Internet content. This can be done by carefully dynamically adapting the mobile device, the wireless link, and the network, for each application — such that we can assure that these dynamic adaptations will work in an optimal way. This is not to say that the network access needs to be application dependent. We assume that access to the network is based on IP over the wireless link. Using this IP-access, the application is able to negotiate for its resources. The adaptation software running in the mobile device utilizes local APIs.

A recent paper [1] shows that we can deliver some multimedia content as background IP-traffic over GPRS. In this scheme, special care must be taken to avoid interference with switched voice services – this results in a roughly 30% under-utilization, so as not to hurt the channel planning of the network operators. A successful strategy, which dynamically shapes the application's demands for Internet content, must address a number of issues:

1. We should avoid situations where we would require unnecessary over-provisioning of bandwidth or other network resources in order for the applications to work.
2. We should maximize the number of users that will be able to use the services in a mobile environment.
3. This will make the applications feasible at an earlier point in time, i.e. *before* network resources are further developed.

4. Such a strategy will save the network provider cost by avoiding investments in infrastructure that might mean *unnecessary* over-provisioning, while providing revenue from these new applications and services.

These issues are particularly important regarding IP-access to wireless networks, where a common assumption has been that we must wait until EDGE or W-CDMA is fully deployed before we can start using these new applications.

3 Proposal

We propose to use agents to represent the different entities in the mobile device, wireless link, and network. The advantage of such an approach is that the agent can behave *intelligently* on a local level. For instance, the agent can transform image content from color to black and white in order to reduce the data that needs to be transmitted, based on the device characteristics *or* the available bandwidth at a given price. In addition, the agents can act intelligently *in concert*, when a certain application demands resources, they can adjust their behavior depending on a *non-local context* (e.g., routing content to (storage) locations, or to where there *are* currently wireless connections with high-bit rates available). Agents may even incorporate machine learning mechanisms to improve their performance with respect to QoS over time.

A common objection to the usage of agents is that solutions involving agents require the global adoption of an agent-specific discovery and negotiation schema to make it work successfully. However, [5,6] show how we can avoid this pitfall by relying on a general signaling protocol (e.g., the Session Initiation Protocol – SIP [10]) for the location of resources and use it to set up a session for the entities to do agent negotiations when applicable.

In the following section, we will show how agents can be applied to the mobile device, wireless link, and network, to optimize the QoS parameters (that were mentioned in section 0) for the application.

3.1 Mobile Device

The application that runs on the mobile device is modeled as an agent. This software is able to use the local APIs to control the speech codecs, length of the sound buffers, choose between available service classes over the wireless link, and even choose between different wireless links.

Furthermore, in certain situations the application may benefit from roaming to another network node. Since the sending party expects a dialog with the receiving party to continue, even if the receiving party goes off-line, the delivery of content is delayed until the receiving party goes on-line again. Utilizing such knowledge about the purpose of the communication can be important in creating intelligent push-services.

In addition, an interesting side point is that advance knowledge of transmissions allows the mobile device to go off-line and hence on standby for longer periods of time, without harming the application, but with dramatic improvement in battery life.

3.2 Wireless Link

Adaptation of header compression profiles [11] and lower-level behavior in the wireless link (e.g., coding, etc.) should not be directly visible to the application. On the other hand, we could introduce an entity in the access network, modeled as an agent, which acting as a proxy will select a suitable adaptation on behalf of the application.

3.3 Network

In a combined GSM and GPRS base station we may include a content-management agent to monitor unused link frames and keep track of which channels are used for switched voice and which channels are used for packet data. This information can be used to increase the utilization of the available bandwidth for IP-connectivity as compared to the relatively static division planned for such base stations.

Furthermore, we can co-locate a SIP-server with the base station. The role of the SIP-server is to locate the user and certain application resources that (as proposed above) should be modeled as agents. Negotiations of sessions are done using SIP, directly between these entities, without involving a SIP-server.

This way, the content-management agent not only helps the base station to increase its throughput of packet data but can also negotiate with the receiver and sender concerning their needs versus available capacity. This provides a basis for developing strategies to adapt the communication in such a way that it fits the end-user's requirements in an optimal way.

The following sections describe two key multimedia applications and illustrate how the proposed functionality can be used to achieve our goals.

4 Streamed Audio

Streaming Audio as broadcasted by Internet radio stations, is not sensitive to delays. In fact, seconds or even minutes worth of buffering can be done. The critical part is to gain acceptance from the user for these types of delays. Recently, the downloading and playback of stored music (e.g., MP3) has become very popular. A forthcoming publication [1] shows how stored music can be forwarded to mobile devices as background traffic in GPRS-enhanced GSM-networks. In addition, user audio (e.g., from dictations or classes with hyperlinks to electronic notes) could be uplinked for later playback. With this type of application, it is understood by the end-user that downloading may take considerable time, but this can be acceptable as long as the download time (typically at night or during the workday) does not exceed the period of time between use.

4.1 Gross vs. Net Content

An interesting calculation is how many unique bits does a radio station produce per day (eliminating the redundant replays of songs, ads, etc.). Now consider an ensemble of radio stations which all play many of the same songs – but have different ads, announcements, and play the songs in a different order — how many unique bits does the ensemble generate? A radio station sending an audio stream at 8 kbps generates roughly 700 Mb during 24 hours. A typical song lasts three minutes. Typically commercials occur four times per hour and last 3 minutes, indicating the station can play 16 songs per hour. This provides room for 384 unique songs but radio station profiling, popularity of certain tunes, and marketing of new music, demand that certain songs appear much more often. Assume, the top ten is played once every two hours, and the next twenty half this rate, and the next ten again at half this rate, hence there will be only 3 random songs per hour. This indicates that the station transmits at best (102 songs * 3 minutes * 8 kbps =) 146 Mb of unique bits. As there are very many radio stations playing popular music network providers can save a lot of bandwidth in the backbone by either coordinating transmissions of identical content or pre-caching content in geographically distributed cells (see further section 0) and perhaps only transmit content identifiers.

4.2 Content Distribution

The agent in the content-server, the base station and the client can co-operate to adapt the communication in different ways. For instance, the agent in the base station may deduce that several end-users have subscribed to the same content and keep local copies in order to reduce traffic in the backbone. It can also multicast this content when possible. Furthermore, this agent can instruct the base station control software to fill unused frames with data, and also predict unused channels for potentially use of delayed the transmission of content. This requires that client agents running in the mobile should be able to deal with paused or even interrupted transmissions, but these are known and solved issues to many FTP-and other clients today.

Based on communication between the agents in the mobile and in the base station, the agent in the base station can note that some of the content does not have to be sent as all the mobiles in the call which would get it - already have this content in their cache.

In light of this, multicasting "ftp" downloads [17], which allow "data holes", should be investigated. The holes can be filled asynchronous to the main download. This way, a user choosing to download an MP3, which is already being downloaded (say halfway) by another user, can hook in to the existing downstream and fetch the rest separately. The actual delivery may even be delayed intentionally until a given number of subscribers are online, or a timeout is exceeded, introducing a notion of content “launch”.

4.3 Caching

We may also think in terms of content delivery networks, for situations where radio resources exist, but the GPRS backbone does not have spare bandwidth. This calls for:

1. Web-server replication (geographically separated, co-located with base stations).
2. Reverse (transparent) cached proxying, possibly implementing hierarchical caches.

4.4 Buffering

A calculation of the necessary storage capacity of large buffers in the mobile device shows that you could go for long periods of time without having more than a very low data-rate high-latency background service. For instance 64 MB holds 7 hours of internet radio quality audio [1] – so at the ~32kbps, which is available during the peak of the day (if you could use all of it) — this amount of memory could be filled in ~16000 seconds (i.e., roughly 4.4 hours).

The capacity of a macrocell (as configured in [1]) can only supply about 64MB of *total transfer* during the peak of the day (unless you want to exceed the 2% call blocking probability). This means you either have to:

1. allocate more capacity to this traffic
2. utilize the unused individual frames within the on-going calls
3. download large portions of the content in hotspots (where you have more available capacity – either because of fewer demands in this cell or because the cell has a higher data rate)
4. download large portions of the content in off peak periods
5. have much larger buffers in the device - so you can pre-load even more content

Dimensioning of buffers is important, it allows us to deploy a Mobile Audio Distribution (MAD) [1] service, where the delay is acceptable as long as you experience continuous audio. It is the user experience or perception, which is important.

Based on the results in section 0, each Internet radio station would send 18.25 Mb/day, assuming some day-to-day coherence the actual amount is *even lower*. This means that during the peak of the day has sufficient spare capacity to support more than three such stations even if these stations had to transmit all their content only during the peak voice hours! All this while using ~32kbps background capacity.

4.5 Agents

Different strategies are possible where companies are able to do directed ads, potentially location or context-aware, paying for the extra cost that this requires in order to have shorter delivery times. Besides adding capacity to either the access network or the device, options 2-3 in section 0 are really the interesting ones where agents are able to help out.

Media files may be streamed from mobile devices to the agent on the access point for subsequent distribution to anyone who desires to listen in. Thus, with GPRS a reporter can collect interviews and broadcast them using at most one channel worth of resources, while simultaneously supporting people listening to MP3-based "stations" on their way home from work. An interesting aspect is when this agent finds superfluous capacity it attempts to pre-order content and trickle it down the wireless link for storage in the mobile.

Furthermore, radio stations often have automated programs, so a network provider could strike a deal with a radio station by buying access to these automatic programming schedules and just multicast the content interspersed with the advertisements to the users at the far end of the network. Exploiting advance knowledge of content programming would save network operators a significant investment while offering radio stations and content providers knowledge about what the user likes, thus offering benefits to all parties!

An important goal for research in the field of mobile computing and communication is to build so-called ‘unconscious’ services: automatic services that do not require user intervention to execute because the components involved are able to make decisions themselves [3,4,5]. An important reason for the use of agents is that they enable us to build such unconscious services. For instance: entering our future homes (where we will have high-bandwidth wireless connectivity between devices) our mobile device will connect to the networked stereo components (where the same information is pre-stored on the server) and hand over the on-going multimedia interaction with the user to it. Before going on a timed standby to save its batteries, the mobile device pre-orders the delivery of a copy of new multimedia content via the server to a content provider, according to the user’s current preferences. Agents represent the components mentioned here. These agents only have to demonstrate reasonable local behavior in order to offer a very useful service in concert. The service relieves the user of consciously having to reload the mobile device with fresh content according to the current preferences.

5 Interactive Voice

The premise that voice will continue to be the most profitable application, which is upheld by many vendors and operators of cellular networks, is false. Voice can be delivered over IP access [9,10] over GPRS [2] even at bit-rates that will be available in the first generation of GPRS. VoIP over wireless can be delivered at 90% efficiency (of the used radio spectrum) as compared to switched voice), when smart header compression is applied [11,12]. Previously, we have shown that 1.2 kbps is all that it takes to deliver voice over IP over wireless [2], which is ~10% of a single GSM voice channel. In further support of the feasibility of VoIP over wireless, DiffServ is a reasonable way of guaranteeing bandwidth for most real-time sensitive applications [1,18] – see further section 0.

Then, we may ask: how long will the statement remain valid that much of the capacity will be used by switched voice, thus relegating anything else to be delivered in the "spare capacity". Delivering switched voice incurs a high investment and maintenance cost for network operators. For wired access, where there is plenty of bandwidth, the cost imbalance has already overthrown the model of switched voice. The "spare capacity" model will remain valid only until the use of VoIP in cellular networks increases sufficient to trigger a wholesale transition to packet-oriented wireless links. But with the experiment that 1.2 kbps is all that it takes to deliver voice [2], this transition cannot be too long after GPRS is introduced. At this average rate you can support at least 26 simultaneous additional VoIP calls in a macro-cell when during busy hours ~32 kbps of spare capacity is available. Assuming business calling patterns this is sufficient to support a population of hundreds of mobile voice users in the area of the cell. Naturally, there will be a QoS reduction caused by the lower speech quality, but end-users will most likely accept this if it means that voice is delivered essentially free in conjunction with other services.

5.1 QoS Considerations

Statically assigned integrated services should be avoided as much as possible, as it implies switched circuit connections over packet networks, whereas differentiated services are crucial to allow e.g. small VoIP packets to cut through the router's FTP and HTTP queues.

Adaptive "loose" integrated services may be necessary, e.g. monitoring the frequency and recurrence of certain high priority packets, capable of preventing issuance of large size packets which would seriously interfere with the probable next high priority packet.

5.2 Base Station

In a recent paper [2] we proposed to use IP directly over the wireless link and thus replace the GSM base station with a router with a radio, thus mobile users will be able to use VoIP for the delivery of interactive voice (see also [7,8]). As compared to the scenario in [1] where switched voice is prioritized, this would provide for much more flexible use of the available bandwidth. The following scenario also holds for the combined GSM- and GPRS-base station, but it is clear that the headroom for dynamic intelligent behavior of the access point diminishes as more channels are allocated/dedicated for switched voice traffic.

It is obvious that such a base station with a content-management agent can adapt the transmission of content in a fair profitable way. As above, the access point can be co-located with a SIP-server.

5.3 Agents

With respect to VoIP, similar scenarios as mentioned in section 0 apply. However, as VoIP puts other requirements on the mobile device, wireless link, and network, agents can help to further adapt the characteristics of these to the current communication context. For instance, if the available bit rate drops, the agent in the mobile device can renegotiate the codec that is to be used, it may also contact the agent in the base station to temporarily halt the transmission of other content that is not sensitive to delays. In case of packet loss, the agent can suggest a change in header compression profiles [11]. The agent in the mobile device could also negotiate with agents located in other base stations for better conditions and more favorable price per bandwidth and cause the communication to be handed over to one of them. Other scenarios may take in to consideration the context of the user (e.g., when the mobile device contains sensors), affecting the content of the voice content. For instance, the content provider could transmit directional audio for augmented reality applications.

6 Conclusions

In telephony applications the load per subscriber during the peak hour corresponds to 3 minutes / 60 minutes = 0.05 Erlang, where a blocking probability of 1% can be tolerated. Assuming we can transmit 26 simultaneously VoIP sessions (see section 0), means that we can serve ~500 voice users in a cell just using spare capacity during peak hours. On the other hand we can make reasonable assumptions about user behavior. If users are talking they will be using other services less (e.g., browsing the web or downloading MP3 files). This means a user agent running with the application in the mobile device can mediate between the two applications that were mentioned in this paper. VoIP applications could convey information on speech activity, similar to 'push to talk' by monitoring input to the speech buffers and thereby signal to the multimedia download application to temporarily pause transmissions [14]. This can also be done by statistical prediction and a 'back off' mechanism.

In addition, for streamed audio applications a much higher blocking probability can be tolerated as long as we can show that we have a throughput that ensures that the content is delivered within a bounded maximum period of time (see also section 0). Above we showed that an agent in the mobile device can mediate on a local level between VoIP and streamed audio download applications. Similarly, the content-management agent can make decisions about the fair distribution of capacity among users of these two applications within the scope of a GPRS/GSM base station through a dialog with the SIP-redirect server. Thereby, the agent has knowledge of ongoing sessions and as mentioned above, can predict traffic and thus assist in planning the transmission of additional Internet content.

The advantage with a agent in the base station is that it can observe the link and utilize the pre-stored bits (MP3, etc.) it has to fill voids in the utilization link - you can have high link utilization but still let voice packets through on time. The same argument regarding monitoring of communication content can be applied to the agent in the mobile device. Speaker dependent speech recognition can be applied using the increased computing capability that is available in mobile devices. A new device first learns "his master's voice". Therefore, voice can not only be surpressed to the point that we send ASCII text "annotated" with voice inflections (consequently causing a further reduction in consumed bandwidth for voice). An agent monitoring content in the mobile device will also be able to alter the mode of communication based upon what is actually being said, which is very useful in applications.

The implications to wireless business models are far-reaching. With GPRS, we have seen that customers get direct access to Internet content. As a consequence, third parties (meaning basically anybody — from private persons, local organizations, to radio- and television broadcasting companies) can provide content to end-users. This means that the network operator has no unique role, unlike the case with WAP gateways. Instead the network operator must seek a new role besides providing competitive price per unit of bandwidth for access to Internet, by offering intelligent support for smart delivery of multimedia content to mobile users. This means providing added value to end-users. At the same time, hosting of Internet content and smart transmission to mobile users creates a business opportunity for network operators to offer these services to content providers. Network operators offering agent hosting can also sell their end-users as a potential audience to those who want to send commercials.

In this paper, we proposed to use agents to represent the different entities in the mobile device, wireless link, and network (that can either behave *intelligently* on a local level *or* act intelligently *in concert*, based on a *non-local context*). By numerous examples we have demonstrated how this approach can be successfully applied to dynamically shape the applications and Internet content, in order to maximize the delivery of multimedia content to mobile users, by taking into account the context of the users and the conditions of the communication. In addition, the transferal of switched voice to VoIP further frees resources for the dynamic shaping of applications and our conclusion is that this transition will take place not too long after GPRS is introduced.

In conclusion, we have demonstrated by our approach of smart delivery of multimedia content in wireless that we can:

1. avoid situations where we would require unnecessary over-provisioning of bandwidth or other network resources in order for the applications to work.
2. maximize the number of users that will be able to use the services in a mobile environment.
3. make the applications feasible at an earlier point in time, i.e. before network resources are further developed, and
4. that such a strategy will save the network provider cost by avoiding investments in infrastructure that might mean unnecessary over-provisioning, while providing revenue from these new applications and services.

7 Future Work

Future work will look at the specifics of the interaction between the SIP-server, the content management agent in the base station, the agent located at the content provider, and the agent in the mobile device. Furthermore we will study how the agent in the mobile device can interact with communication resources, in particular those governing VoIP QoS, e.g. DiffServ and mobility. Another issue that needs to be addressed concerns how the content-management agent interacts with the radio resources in the base station.

8 References

1 P. Lindtorp, “Utilizing Spare Capacity in Radio Access Networks”, Master Thesis, School of Electrical Engineering and Information Technology, Royal Institute of Technology, Sweden, December 1999.

2 T. Kanter, C. Olrog, G. Maguire, "VoIP over Wireless for Mobile Multimedia Applications", Personal Computing and Communication Workshop, November 1999.

3 T. Kanter, C. Frisk, H. Gustafsson – “Context-Aware Personal Communication for Teleliving”, Personal Technologies (vol. 2 issue 4, 1998: p. 255 - 261).

4 T. Kanter and H. Gustafsson, “VoIP in Context-Aware Communication Spaces”, Proceedings of International Symposium on Handheld and Ubiquitous Computing (HUC 99), Oct. 1999.

5 T. Kanter and H. Gustafsson, “Active Context Memory for Service Instantiation in Mobile Computing Applications,” Proceedings of the Sixth IEEE International Workshop on Mobile Multimedia Communications (MoMuC'99), Nov. 1999, p. 179-183.

6 T. Kanter, “Adaptive Personal Mobile Communication”, forthcoming Licentiate Thesis, Dept. of Teleinformatics, Royal Institute of Technology, Sweden.

7 C. Olrog – “GSM SoftModem on Linux, a Direct Radio Link Protocol Interface”, Master Thesis, Dept. of TeleInformatics, Royal Institute of Technology, Sweden, April 1999.

8 T. Turletti, H. Bentzen and D.L. Tennenhouse, ``Towards the Software Realization of a GSM Base Station", IEEE/JSAC, Special Issue on Software Radios, Vol. 17, No. 4, pp. 603-612, April, 1999.

9 ITU-T Recommendation H.323 v2 (1998) Packet Based Multimedia Communication Systems.

10 M. Handley, H. Schulzrinne, E. Schooler, J. Rosenberg - RFC 2543 on SIP: Session Initiation Protocol, IETF/Network Working Group – March 1999.

11 L-E. Jonsson, M. Degermark, H. Hannu, K. Svanbro, “RObust Checksum-based header COmpression” (ROCCO), IETF Network Working Group Draft, September 1999.

12 M. Engan, S. Casner, C. Bormann - RFC 2509 on IP Header Compression over PPP, IETF/Network Working Group, February 1999.

13 J. Mitola III, Cognitive Radio, Licentiate Thesis, Dept. of Teleinformatics, Royal Institute of Technology, Sweden, Sept. 1999.

14 J. Mitola III, “Cognitive Radio for Flexible Mobile Multimedia Communications,” Proceedings of the Sixth IEEE International Workshop on Mobile Multimedia Communications (MoMuC'99), Nov. 1999, p. 3-10.

15 V. G. Bose, "The Impact of Software Radio on Wireless Networking," Mobile Computing and Communications Review, Volume 3, No. 1, January 1999.

16 News articles on www.planetunreal.com, "UT Over Wireless" and “UT Over GPRS”, 11/19/99.

17 J. Ioannidis, G. Q. Maguire Jr., “The Coherent File Distribution Protocol,” RCF 1235, IETF Network Working Group.

18 S. Black, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, “An Architecture for Differentiated Services,” RCF 2475, IETF Network Working Group

IPv6 : The Solution for Future Universal Networks

Sathya Rao

Telscom AG, Sandrainstr. 17
3007 Bern, Switzerland

Abstract. The communication networks and services are changing rapidly. The conventional circuit and packet switched networks are being replaced by next generation networks, primarily based on Internet Protocol. The rapid growth of web based services has lead to the explosive growth of the internet. However, the current internet protocol (IPv4), which is the backbone of transmission control protocol (TCP/IP) networking, is rapidly becoming obsolete, with the inherent problems related with limited address space, security and QoS features. The new protocol IPv6 has been developed to overcome all these problems and to provide solutions for the next generation networks. This paper addresses the features of IPv6 Protocol, the status of standardisation, and various activities around the world.

1 Introduction

Europe's leadership in Internet technology and provision of user access should be based on an offering with unlimited address space, quality and security, properties the current Internet does not cater for. Europe should foster a unique leadership strategy in promoting the next generation networks based on the new Internet Protocol version 6 (IPv6) protocol in order to promote pan-European E-commerce, offering customer protection and benefits in terms of security and quality as services converge to run over IP. Such a Euro-IPv6 network will place Europe into a position of strength in comparison to the US with respect to New Internet technology.

The deployment of IPv6 requires a good spread of diverse technologies and the support of national Internet Service Providers across the whole European community. Expertise in these new technologies, which overcome the limitations of the current IPv4-based Internet, cannot be found in just a couple of European countries. The skills required lie in the areas of seamless deployment of IPv6 into a large existing IPv4 base, and provision of quality of service and security at host and gateway/router level.

The Internet has doubled in size every year since 1988. There are over 44 million hosts on the Internet and an estimated 200 million users world wide. By 2006, the Internet is likely to exceed the size of the global telephony network, should IP telephony have not replaced the existing telephony network by then. Moreover, tens of millions of Internet-enabled appliances will have joined.

C.G. Omidyar (Ed.): MWCN 2000, LNCS 1818, pp. 82-91, 2000.

The Telecom industry (manufacturers and operators) need to build strategies to cater for the mobile information society, deploying brand new products and services such as wireless Internet devices, Internet cell phones and personal digital assistants which will emerge to become the new telecommunications tool of the next decade. The mobile information society will need to deploy for this purpose IPv6 as a robust Internet foundation.

This strategy will get the European leadership entrenched in every aspect of the European industry in view of the explosion of the E-business and E-entertainment in Europe. It is estimated that commerce on the network (E-commerce) will reach somewhere between 1.7 T$ and 3.0T $ by 2003. That is only three years from now (but a long time in Internet years). Secure E-business and European privacy should be advocated and implemented at the network layer not just at an application layer. The security feature is built-into the IPv6 protocol to solve the issue, which is one of the major weakness of the current internet protocol.

2 Applications with IPv6 at Driving Seat

E-commerce will be a new driving force for new economies, creating new business sectors and new jobs across Europe. This new economy needs a robust platform to guarantee its success. The E-business and E-shopper should be able to gain the necessary confidence that they are doing business in a safe environment, which is not the case today. European E-commerce could back-fire in the mid to long term without adequate customer protection. E-commerce will also create the need for more address space and this new need is a healthy sign of the growth of the Internet in Europe, but then this growth has to be supported by guaranteed IP address space which is not available today.

The Internet is proving to be one of the most powerful amplifiers of speech ever invented. It offers a global megaphone for voices that might otherwise be heard only feebly, if at all. It invites and facilitates multiple points of view and dialogue in ways not possible through traditional, one-way mass media.

The Internet can facilitate democratic practices in unexpected ways. The proxy voting for stock shareholders is now commonly supported on the Internet. Perhaps we can find additional ways in which to simplify and expand the voting franchise in other domains, including the political, as access to the Internet increases.

The Internet is becoming the repository of all we have accomplished as a society. It is becoming a kind of disorganized Boswell of the human spirit. Shared databases on the Internet are acting to accelerate the pace of research progress, thanks to online access to commonly accessible repositories.

3 Ongoing Activities

The US government is funding the 6REN/6TAP testing native IPv6 as well as Internet2, a project that tests the impact of QoS and higher bandwidth on the current Internet. Internet2 has subscribed now to IPv6 as the result of their tests that higher bandwidth is not the only solution but a smarter packet is needed to achieve a better quality of service.

The Japanese government is funding the Wide Project which is a copy of the 6REN/6TAP initiative to take Japan into leadership in the New Internet. The Japanese government is pushing for active IETF involvement in order to secure RFC adoption or early RFC influence.

The IPv6 Forum has been formed recently to promote wide adoption of IPv6 specifications in developing next generation network products and services. The IPv6 Internet Initiative is a key milestone for a range of products and services under definition within the mobile information society platform. European Concepts based on GPRS, UMTS and 3G products and services depend dramatically on the deployment of IPv6. The convergence is an opportunity for the European switch manufacturers to take leadership into the New Internet and define Core Switch/Routers.

Within the European Union 5th framework 'Information Society Technologies' framework, a project named 6INIT has been started to promote the deployment of IPv6 networks and services, in collaboration with Japanese and Canadian partners.

Eurescom has initiated a project (P702 : Internet Protocol Version 6 - new opportunities for the European PNOs) to investigate the usage of IPv6 networks to replace the conventional networks for delivering conventional and future public services. Eurescom has also initiated a new project (P1009) to study the deployment and transition strategies for services on top of IPv6, e.g. Mobility, QoS support, Multicast Implementation, Network configuration and management.

4 Standards Status

4.1 The Internet Engineering Task Force and IPnG

The IETF (Internet Engineering Task Force) is very active in promoting IPv6 standards, through their Request for Comments (RFC) documents which are generally adopted as standards for implementation. IETF has constituted a special group IPnG (IP next generation) to promote IPv6 activity.

The current version (4) of the Internet Protocol (IPv4) uses node addresses that are allocated from a 32-bit space This 32 bit address space is further classsified to provision Class A, B and C ranges, which constituted network part of 8, 16 or 24 bit, with corresponding host part of 24, 16 or 8 bit, depending on the number of expected hosts on a given network. This led to inefficient use of the 2^{32} possible addresses, since many ´important` organisations automatically asked for class A or B addresses using up 2^{24}or 2^{16} addresses at each single assignment, even when they often only had several host computers, or had many subnets with several computers on each. A second problem was that addresses were rarely re-claimed after they were no longer in use.

The terms of reference for the working group is maintain all good features of current protocol specifications, and enhance the features to guarantee smooth transition to next generation networks. The IPv4 protocol is simple, binds multiple protocols, simple management, but limited with scalaibility in terms of address space, topological flexibility, Quality of service support, security, etc..

IPv6 is planned to support very high speed (Gbps), range of subnets, low information loss and will function independent of media (terrestrial, mobile, radio and satellite), provides auto configuration possibility, high security for business applications, application specific QoS, and multicast addressing facility. IPv6 will be also backward compatible to work with the current IPv4 protocol (through tunnelling mechanism).

4.2 IPv6 Features

The **next generation networks** based on IPv6 will provide:

- 128 bit wide address space to cover all possible appliances connectivity
- Differentiated Services in terms of quality (bandwidth guarantee and transit delays for real time flows).
- Security in terms of access point authentication, message integrity and privacy.
- Auto-configuration and reconfiguration capabilities allowing easy modification of network architectures.
- Management facilities allowing the setting up of on-demand services and providing ISPs with accounting capacities.
- Wide range of applications and services.
- Mobile host capabilities allowing provision of transparent access whatever the physical access used, supporting the evolving UMTS capabilities, will be the issue of co-operation between the mobile IP related projects (e.g. WINE).

4.3 Activities in the IETF

Within the IETF now, detailed work on IPv6 specification is persued. Changes to routing protocols, transport protocols (the pseudo-header checksum in TCP and UDP) and applications that reference IP addresses (particularly DNS, but also FTP) have been specified. More subtle wok in routing (beyond OSPF and RIP v6 changes) needs to be done, and more especially, the impact on RSVP and on multicast routing and Mobile IP routing as well as RTP/SDP and MPLS needs a lot of work to see what the real benefits may be.

The critical missing piece in IPv6 is a deployment plan that includes seamless interworking with IPv4, but provides clear benefits to a site to migrate. Three possible ways this may happen are: VPNs, Satellite IP (DBS) and large scale PDA/GSM mobile phone integration by a provider and vendor. The importance of the seamless interworking is because the Internet is now far too large to envisage event the switchover that occurred in going from previous NCP to IP in 1980; and that switchover was even painful then - with a few hundred hosts rather than tens of millions. Moreover, there is clear reluctance by commercial vendors to invest into extensive upgrade; it must be made worthwhile by the quality of the benefits provided.

4.4 Available Implementations, Products, and Services

We can divide implementations into host and router side code. In the router side, most of the major vendors have at least beta products for the basic IPv6, although its not clear if their routing protocols changes they still interwork with the legacy IPv4 networks. On the host side, major operating systems such as Windows 2000 will have IPv6 in, although to date, only the research part of Microsoft have released a Windows implementation. In a recent IPv6 Summit Microsoft announced their commitments to IPv6 and officially released the IPv6 protocol stack. Similarly CISCO also announced their commitment to the IPv6 networking, which provided early boost the next generation networks world. For Unix systems, there are releases for most major flavours, although many are very early code, and have a number of shortcomings. The best systems are the public domain offerings for FreeBSD and Linux, including DNS for IPv6 and other important infrastructural tools.

The implementations have many of the components that have been defined - but not all. For example, strong security is mandatory in IPv6; political considerations related to export controls have made it very difficult to have exportable implementations which meet the Standards. Moreover, some of the key components, like IP Multicast, Quality of Service facilities and Public Key Infrastructure are not yet fully standardised.

4.5 Address Space in Reality

In practice, the IPv4 address space has lasted longer than expected due to two technologies: Classless Inter-Domain Routing (CIDR), and Network Address Translation (NAT). This has reduced the urgency to move to IPv6.

CIDR is a generalisation of the class based address assignment that was originally devised for IPv4. Nowadays, the address assignment authorities (and they are devolved to regions of the Internet geographically now) assign IPv4 (and IPnG (or IPv6 as it is known to some)) addresses hierarchically, together with masks. The mask determines how many bits of the address are network and how many are host, and the mask can be different at different places in the network topology - this allows the address + mask to be treated like a variable length prefix. The forwarding decision that used to be made by routers, based on simple best-next-hop, is no longer a simple lookup; it consists of a longest-match procedure. Routing protocols no longer exchange lists of network numbers to build upon a network map, but now exchange addresses + masks, to allow this hierarchical address space management to work. A secondary, but very important side effect of this is that the routing tables can now be summarised; typically, a country might be assigned a short mask, and within the country, each region, longer masks. This allows each router nearer to each local region to contain small number of (even just 1 per interface) entries.

The questions of the technical and political feasibility of address-space deployment are relatively separate. One advantage, for example, of the larger address space is that mobile users can keep their same low-order addresses, while the mobile network operator controls only the high-order bits. However when there is a real conflict between technical and commercial pressures, the solution is less clear. It would be possible to carve out a complete set of numbers and address for IP-telephony; early attempts to do this in conjunction with the ITU telephony groups have failed so far. One suspects that this is partly because attempts to make IP numbering as universal and well-structured as the telephone numbering is seen as a very real threat by the PNOs to their telephone revenue.

NATs are another technique for containing the growth of address use, but are based in two other important requirements: to avoid having to renumber hosts, and to provide some network security. Many sites had assigned IP addresses in isolation from the Internet, and had used addresses already in use in the world-wide Internet. To avoid the problem of re-numbering all their hosts and routers, such sites developed a technology to translate ´on-the-fly` the addresses of source hosts within IP packets being forwarded through firewall routers. This was done only for hosts which wished to communicate with the ´outside world`, changing the addresses in packets to and from them, from the internal non-unique ones, to ones drawn (dynamically assigned) from a small pool of legitimate public addresses. This works well in typical large corporate networks as few hosts communicate with the outside world at any one time (principal of locality!). Often, the dynamic assignment in the firewall router was controlled by an application level authentication protocol (e.g., based on an RPC mechanism, or even just a telnet/login user-name + password to the firewall router). This meant that the NAT acts as a quite effective barrier to outside hosts accessing internal hosts (even if packets could get in by pure random luck, the responses would not get out...).

4.6 IPv6 Roadblocks

Many of the improvements promised for IPv6 have been specified in a simplified way for IPv4; the specifications are often less powerful than is possible with IPv6, but would provide enhanced services. It has turned out to be very difficult to organise even the IPv4 deployments of IPSEC, QoS and Mobile IP. This is partly from lack of motivation by the many smaller ISPs, but it is also because of the need to orchestrate the introduction of the services. There is often little advantage in introducing such services in an isolated part of the system. Moreover, in some cases the specifications are not really complete, or it is felt that their viability yet to be demonstrated by the R&D community. Here there is the serious problem that small-scale deployment is no feasibility demonstration; large-scale deployment is often beyond the means of the R&D community, and is not really supported by those providing the research networks.

It is probable that these improvements will come more with IPv6 than earlier, because the whole of the transition to IPv6 must be managed to some extent in any case. This will encourage the establishment of larger testbeds, at least by the bigger ISPs and PNOs. There are some substantial test-beds already; who are active on the 6Bone, which is a set of European sites who are experimenting with IPv6 in an encapsulated form. During the last summit NTT (from Japan) announced the world's first ISP to support IPv6.

5 Evolution Scenarios

The Internet engineering community is promoting a new version of the IPv6 as the answer to the address shortage predicted for the current Version 4. IPv6 offers enough addresses that every computer, cell phone and set-top box can be hooked up to the 'Net. However, migrating a large network to IPv6 is so difficult that few organizations have committed to it.

Theoretically, Version 4 could support up to 4.2 billion devices, but the allocation of those addresses has not been very efficient. An attempt has been made to increase the efficiency with interdomain routing and allocation rules that go along with it. But the side effect of those rules is the proliferation of network address translation [NAT] boxes, which take a single Internet address and multiplex it among a bunch of different devices. It's a fairly ugly process from an architectural point of view, although it turns out to be very effective, and a lot of people are relying on it. But because NAT intervenes at the IP address level, it has some consequences for end-to-end security and integrity of the traffic

The key transition objective is to allow IPv6 and IPv4 hosts to Interoperate. A second objective is to allow IPv6 hosts and routers to be deployed in the Internet in a highly diffuse and incremental fashion, with few interdependencies. A third objective is that the transition should be as easy as possible for end-users, system administrators, and network operators to understand and carry out.

Probably the most straightforward way to introduce IPv6-capable nodes is a dual stack approach, where IPv6 nodes also have a complete IPv4 implementation as well. Such a node, referred to as IPv6/IPv4 node in [RFC 1993], thus has the ability to send and receive both IPv4 and IPv6 packets. When interoperating with an IPv4 node, an IPv6/IPv4 node can use IPv4 packets; when interoperating with an IPv6 node, it can speak IPv6. IPv6/IPv4 nodes must have both IPv6 and IPv4 addresses. They must furthermore be able to determine whether another node is IPv6-capable or IPv4-only. This problem can be solved using the DNS, which can return an IPv6 address if the node name being resolved is IPv6 capable, or otherwise return an IPv4 address. Of course, if the node issuing the DNS request in only IPv4 capable, the DNS returns only an IPv4 address.

In the dual stack approach, if either the sender or the receiver is only IPv4-capable, IPv4 packets must be used. As a result, it is possible that two IPv6-capable nodes can end, in essence, sending IPv4 packets to each other. This is illustrated in Figure 1.

The target scenario is to have an IPv6 backbone network, which can also provide seamless interconnectivity with legacy IPv4 network. The typical network scenario with IPv6 backbone is shown in the Figure 2.

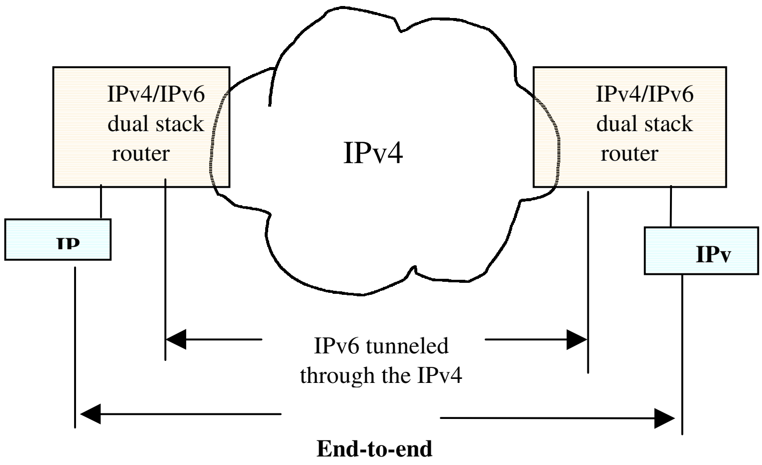


Figure1: A dual stack approach

6 Mobility and Internet

Europe is very strong in mobile networks deployment and usage. The internet access through mobile terminals and having an interactive data communication is a priority area in Europe. The fist such applications already being introduced based on WAP.

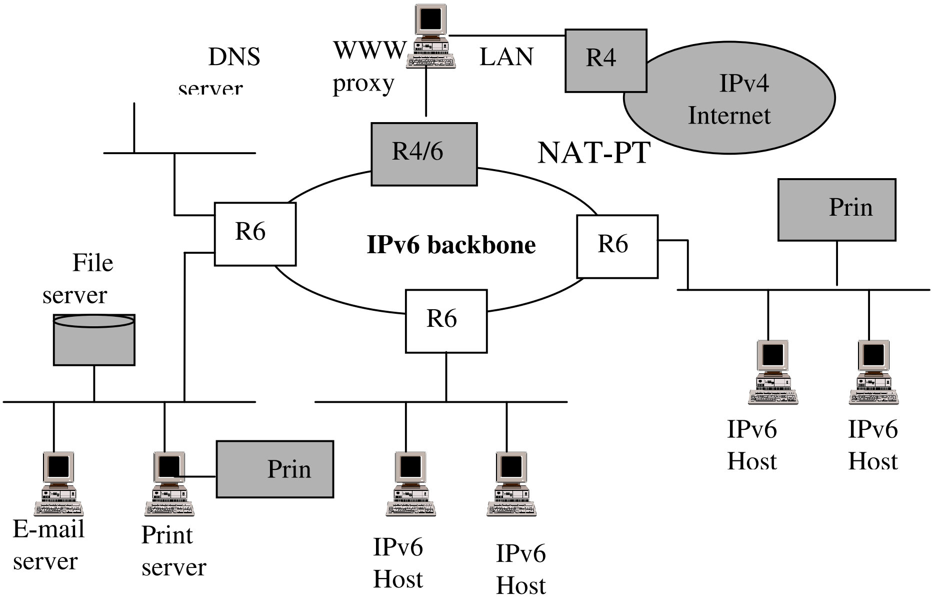


Figure 2: IPv4 and IPv6 network interworking

The WAP based communications are slow due to bit oriented non-efficient WAP protocol, though it provides the transition step for introducing mobile internet to the market. Third generation networks deployment plans to implement UMTS services are already in advanced stage of realisation. To progress the evolution and to enhance both mobility and internet features, the new initiative has been put in place called Third Generation Partnership Project (3GPP). The 3GPP is a global standardization initiative that was created just over a year ago, in December 1998, to produce technical specifications for Third Generation Mobile System based on the evolved GSM core networks and a new radio interface (UTRA). Major important steps have been achieved since then, such as the approval of Release '99 specifications in December 1999. The 3GPP work plan for the year 2000 includes Internet Protocol (IP) based communications. It is expected that as mobile phones gain access to Internet services, there will be an unprecedented growth in the demand of new Internet addresses as well as easier administration and tighter security. Convergence of Internet and Mobile Telecommunications move a step closer since IPv6 Forum has joined the 3GPP as a Market Representation Partner.

7 Conclusions

The specifications for a Next Generations Internet are largely complete from a technical viewpoint. There are still some loose ends, but they are not very significant. The large-scale deployment of many of the newer features over the current Internet is proving difficult, and will probably never happen on a large scale. Implementations of the basic feature sets are available in research prototypes, and starting to become available in commercial offerings. Implementations of advanced features are becoming available in research prototypes, but still require substantial experimentation and refinement. However, there are starting to be a number of substantial research networks on which the implementations are being deployed for R&D purposes. The general commitment to large-scale commercial deployment, and the time-scales over which this could be achieved, are still under discussion, though the recent announcements from the major vendors has brought the time scales to near short term plans.

Performance Measurement Methodologies and Quality of Service Evaluation in VoIP and Desktop Videoconferencing Networks

Henri TOBIET[1] and Pascal LORENZ[2]

[1] NMG TELECOMS – Network Management Group – Telecoms Solutions
20e, rue Salomon Grumbach, BP 2087, 68059 Mulhouse Cedex, France
h.tobiet@nmg.fr
[2] University of Colmar - IUT / GTR
34 rue du Grillenbreit - 68008 Colmar, France
lorenz@colmar.uha.fr

Abstract. The present paper relates to performance and quality of service evaluation for VoIP (Voice over IP) and desktop videoconferencing services in IP networks. Simulation-based performance measurements consist in the generation of performance statistics obtained by measurements realized by simulation on the subscriber interface. They include detailed measurement of call quality, call set-up quality and availability. The tests are accomplished by emission of non-disturbing additional test traffic. The real-time QoS monitoring is based on non-intrusive analysis of real call parameters. The advantage of this method is the exhaustiveness of the analyzed call. The QoS measurements are mainly service availability and call set-up quality.

1 Introduction

QoS defined in CCITT Recommendation E.800 may be considered as the generic definition reproduced below: *"The collective effect of service performance which determine the degree of satisfaction of a user of the service".*
Quality of Service on the LAN represents a major challenge, not so much during the predictable processes of compressing the voice streams and splitting them into packets which are mathematically predictable, the real challenge is of sharing the connectionless transmission media with other users in a predictable and quantifiable way. As soon as voice/video traffic reaches the IP network, it must compete with electronic mail traffic, database applications and file transfers [13], [17], [18].

QoS evaluation methodologies are based on the previous studies performed in European Projects, such as QOSMIC (QoS Methodologies and tools for Integrated Communications). The resulting QoS will be dependent on the performance of the physical, AAL (ATM Adaptation Layer) and transport layers involved in the protocol stack implementation [1], [10].

Work is currently on-going in modelization of the TCP/IP protocol layer, in terms of performance aspects (delays, etc...). The resulting model will then be applied to

C.G. Omidyar (Ed.): MWCN 2000, LNCS 1818, pp. 92-107, 2000.

convert network layer performance (ATM, AAL) to transport layer performance, according to the selected services: telephony over IP, file transfer over IP and videoconferencing over IP [12], [14].

1.1 QoS Characterization

The used TCP/IP model can be described as follows:

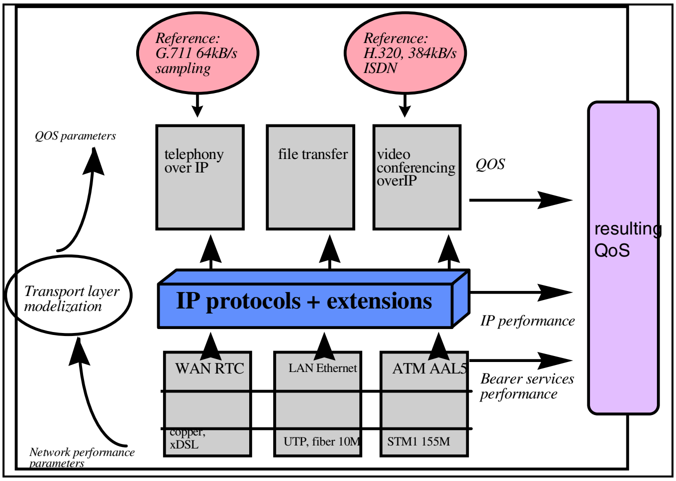

Reference: G.711 64kB/s sampling
Reference: H.320, 384kB/s ISDN
QOS parameters
telephony over IP
file transfer
video conferencing overIP
QOS
resulting QoS
Transport layer modelization
IP protocols + extensions
IP performance
WAN RTC
LAN Ethernet
ATM AAL5
Bearer services performance
copper, xDSL
UTP, fiber 10M
STM1 155M
Network performance parameters

Fig. 1 : QoS protocol oriented approach

The main performance parameter to be measured is the Round Trip Time (RTT) but the performance evaluation will also concentrate on the following aspects: connections opening and closing mechanisms, data transfer mechanisms, addressing, parameters negotiation, congestion control and errors control.

Part of our studies concentrated on end-to-end QoS characterization. End-to-end QoS in a videoconferencing system is characterized under two broad headings:

- call set-up quality and
- call quality.

Call set-up quality is mainly characterized by the call set up time, i.e. the time elapsed from the end of the user interface command by the caller (keypad dialing, email alias typing, etc) to the receipt by the caller of a meaningful tone. ITU-T Recommendation E.600 provides more information on the definition of post dialing delay. Call set-up time is perceived by the user as the responsiveness of the service. Other factors such as ease of use contribute also to the user experience. The first of these factors is objective, the second is subjective [16].

Within the broad category of call quality two major factors contribute to the overall QoS experience of the user of the videoconferencing system:

- the end-to-end delay which impacts the interactivity of a conversation and
- the end-to-end video and speech quality.

The following factors contribute to the overall call-set up time:

- IP access network set up delays (these would include transport layer set up-times, modem training times and log on times at the ISP Gateway),
- signaling delays across the IP backbone,
- call set-up delays within the gatekeeper(s),
- access times and call processing delays to back-end services, such as directory services or authentication services,
- call set-up delays within the gateway,
- call set up times in the network(s).

The end-to-end delays are influenced by IP terminal buffering delays, H.323 packetization/buffering delays, codec delays and network transmission/propagation delays. The end-to-end audio/video quality depends on input and output devices, analogue/digital and digital/analogue circuit noise, video and audio coding distortion, effects of bandwidth limitation in the IP network.

The QoS issues associated with the IP terminal are the choice of codecs used in the terminal, the performance of the codec to various types of network degradation, the signal processing delays, the call processing delays, the number of frames per packet, the processing delays associated with security issues, the design of jitter buffers, the delays through the audio or digital media paths and the performance of echo-canceling devices [7], [8], [11].

1.2 QoS Issues Associated with LAN Access

In this configuration the access layer is limited to the Network Interface Card (NIC) used within the IP terminal. Though the LAN has ample bandwidth for transmission of coded speech/video, a fundamental issue frequently encountered is contention for shared media.

At any time, other (non audio) endpoints on the LAN, may flood the LAN and consume all the available bandwidth. This problem can only be avoided if there are mechanisms to manage and police the use of bandwidth (both for real-time use and best-effort use). The Subnet Bandwidth Manager (SBM) and RSVP (IETF RFC 2205) are intended to provide this capability.

The factors affecting QoS in this scenario are the transmission delays through NIC and the jitter in data buffers associated with the NIC. It is anticipated that these parameters will in general be well controlled and specification of upper bounds on these parameters should present few difficulties.

1.3 QoS Associated with PSTN Modem Accesses

In this type of access, modems are used to establish a digital channel between the videoconferencing terminal and the IP network. The factors affecting QoS in this configuration are the:

- modem bit rate,
- modem transmission overheads,

- throughput delay in modem and at ISP site,
- jitter within client modem, ISP modem and buffers,
- PSTN set-up time,
- modem connection set-up time,
- ISP logon & set-up time and
- error rate on PSTN link.

1.4 QoS Associated with ISDN Accesses

ISDN access uses a set bandwidth for the communication channel (16 kbit/s for the D channel, 64 kbit/s for a B channel). Aggregation of n*B channels to provide a 384 kbit/s channel provides a means of using video codecs even with normal RTP/UDP/TCP/IP overheads. The factors affecting QoS in this scenario are the:

- use of PPP/IP/UDP/RTP header compression on access link,
- throughput delay in ISDN terminal adapter and at ISP site,
- jitter within ISDN terminal adapter and ISP network interface buffers,
- ISDN set-up time and
- ISP Logon and session set-up time.

1.5 QoS Issues Associated with xDSL Accesses

xDSL access allows the use of various sizes of bandwidth, up to tens of Mbit/s, depending on application and the DSL technique used (e.g. ADSL, VDSL). IP access may use in general a mediation transport layer (i.e. ATM) or may be mapped directly into the xDSL frame (not standardized yet). The factors affecting QoS in this scenario are the:

- xDSL modem available bit rate (due to line condition and specific application),
- use of PPP/IP/UDP/RTP Header Compression on access link,
- throughput delay in xDSL modem (fast or interleaved) at ISP site,
- jitter within client modem, ISP modem and adaptation buffers,
- xDSL set-up time (e.g. when using Dynamic Power Save in VDSL application),
- ISP Logon and session set up time and
- error rate on access link.

2 Typical QoS Measurement Campaigns

2.1 Intrusive Measurement Tool = QoS Simulation Platform

The main functions of the requested tool are :

- capture and analysis of the received traffic at ATM, IP and application level, taking into account the different network architectures (IPoATM, LANE/MPOA, MPLS, etc.) and their characteristics,
- performance measurement at different network layers, focusing on relevant metrics such as: throughput, one-way delay, delay variation, packet loss, etc as referred in the RFCs 1944 and 2330,

- objective QoS evaluation, by taking into account the relevant performance parameters and network characteristics (e.g. scheduling techniques like WFQ or CBQ),
- generation of calibrated multimedia traffic patterns made of RSVP messages,
- traffic allocation to QoS Service Classes and CoS contract verification of IP streams based on Tspec parameters using a token bucket algorithm as referred in the RFC 2215. This verification would allow these streams to fit a DiffServ or IntServ network requirements,
- comparison with subjective data relating to the quality as seen by the end-users.

Iterative experiments will allow the tool to increase its knowledge in mapping objective and subjective QoS, to be able to allocate automatically the received traffic to the corresponding Service Class, only by measuring the adequate network performance. The simulation tool will enhance its QoS evaluation process by implementing innovative self-learning techniques.

2.2 Non- intrusive Measurement Tool = QoS Monitoring System

The monitoring tool (QoS Probes + Supervision System) will have following functions :

- processing the self-learning algorithms preliminary established by using the Simulation Tool in similar network environment,
- accessing the network in real-time, without disturbing the current traffic,
- traffic allocation to QoS Service Classes,
- results supervision via data collection by a supervision system, which will be designed to be connected, via TMN interfaces, to the network administration system,
- supervision system design taking into account entities from DiffServ architecture like Bandwidth Brokers (BB) or protocols like Common Open Policy Service (COPS).

2.3 End-to-End QoS Experiments

Experimental services will be based on multimedia applications. Voice over IP quality will be determined according to G.711 64kB/S characteristics. Video over IP will refer to H.320 ISDN video-conferencing running at 384KB/S. QoS expertise will take into account most types of coding and compression algorithms (G.723, H.323, M/J-PEG.).

Performance will be measured at standardized access points :

- PCM n*64KB/S,
- ISDN, basic and primary accesses,
- PDH, SDH (STM1-STM4),
- xDSL subscriber loop accesses (probably ADSL),
- Cable-modem accesses,
- ATM (PDH 34MB/S, SDH STM1-STM4).

The evaluation process will make use of self-learning techniques, allowing statistical approach of QoS parameters. QoS expertise will take into account real-time enhancements of transport protocols (UDP/IP, RSVP, RTP, RTCP, IPV6...). As QoS management involves all terminal and network related aspects, measurement and monitoring will cover most of them. These include Tspec definition, Adspec definition if RSVP is to be used, RSVP/ATM issues, DiffServ/IntServ interface issues, DiffServ/ATM issues, MPOA/LANE QoS if used, IPv6/ATM, etc... Measurement techniques will be enhanced in order to anticipate performance degradation and to predict QoS evolution [2], [3].

Performance evaluation will be characterized by following aspects:

- connections opening and closing mechanisms,
- data transfer mechanisms,
- addressing, parameters negotiation,
- congestion and errors control,
- bandwidth allocation,
- CoS contract verification.

3 Voice over IP Experiments

3.1 ETSI TIPHON Project Simulation Experiments

The ETSI TIPHON contribution 09TD42 presented a proposal to carry out simulations where different scenarios of end to end speech transmission over IP based networks could be evaluated with respect to speech transmission quality.

In autumn 1998, an investigation according to 09TD42 was performed by Deutsche Telekom Berkom (T-Berkom) where such different scenarios were simulated and subjectively assessed in a well-established listening. The simulation processing contained a couple of speech codecs, packet loss ratios and various kinds of audio frames per IP packet. This TD describes the test methodology, the simulation method, the scenarios and the results of the executed simulation processing.

Furthermore we would like to discuss some interesting results concerning the relationship of speech material (construction, length) packet loss and their influence to the auditory assessment.

3.1.1 Test Methodology

TIPHON WG5 (DTR/TIPHON 05001 V1.2.5, chapter 7.3.2) defined a methodology for testing speech quality in TIPHON compliant networks and terminals. This methodology was taken into account and used as a basis model for the T Berkom simulation processing.

A set of speech signals designed according to ITU-T Rec. P.800 was used as input of the simulation path. The simulation path includes the terminal side (electrical part) and the network itself. The influence of the terminal side was focussed to the speech conversion and IP packet size issue. The influence of the network side was simulated by different packet loss rates. After the simulation the speech samples were recorded and stored in a database.

3.1.2 Simulation Method

For simulation of network influences in the case of packet loss, a common channel model was designed, realized by channel files which describe the network condition with the same time resolution as the source speech sample rate. So the network has a certain condition (good or bad) for every speech sample (every 125 µs) two adjacent network states were considered as statistically independent, because the network speed was assumed to be much higher than the sample rate (8000 samples per second). So for each packet loss rate one channel file was created using a random generator. The length of this channel file was exactly the same as the length of the speech file.

In a further step the speech file, assembled in IP packets, was matched to the channel file. According to the length of the IP packet (10ms, 20ms,...) the channel file was checked every time when a packet was ready to send. That means if the packet size was 10 ms the channel file was checked also every 10 ms if the condition is good or bad. In a bad case the IP packet was lost, otherwise it was further processed. This information (IP packet lost or not) was stored in a description file which was the input of the re-assembler and speech decoder.

3.2 Testing of Speech Quality

There are two methods of testing end-to-end (acoustic to acoustic) speech quality:

- subjective tests involving the opinion of panels of users (see ITU-T Recommendation P.800),
- objective tests including comparison methods against a known reference signal (See ITU-T Recommendation P.861), absolute estimation methods.

Based on ITU-T Recommendation P.561, the measurement of individual parameters followed by the use of a Transmission Rating Model (TRM) to combine the effects of the individual parameters and predict the subjective views of users. The E-model is under consideration for this purpose.

Subjective tests have the advantage of including all parameters and providing a direct subjective view, but they take a long time to perform, are costly and are ill-suited to investigating changes in the values of many parameters because of the large numbers of combinations involved.

Objective comparison methods are described in DEG/STQ00001. Objective tests using the E-Model approach should include the same parameters as in the PSTN world:

- SLR	Sending Loudness Rating;
- RLR	Receiving Loudness Rating;
- OLR	Overall Loudness Rating;
- STMR	Sidetone Masking Rating;
- LSTR	Listener Sidetone Rating;
- Ds	D-Value of Telephone at Send-side;
- Dr	D-Value of Telephone at Receive-side;
- WEPL	Weighted Echo Path Loss;
- qdu	Number of Quantizing Distortion Units;
- Ie	Equipment Impairment Factor (low bit-rate Codecs);
- Nc	Circuit Noise referred to the 0 dBr-point;
- Nfor	Noise Floor at the Receive-side;
- Ps	Room Noise at the Send-side;
- Pr	Room Noise at the Receive-side.

For evaluation of the Ie values for low bit-rate codecs, some objective measurement methods have been developed but commercial measurement systems are not yet available. In addition, specific requirements from the TIPHON system (eg. packet loss) have to be considered in determining Ie.

In conversational situations:

- TELR	Talker Echo Loudness Rating;
- T	Mean one way delay of the echo path; and
- Tr	Roundtrip Delay in a closed 4-wire loop,

need also to be considered.

The performance of TIPHON systems in terms of TIPHON speech quality classes may also be measured between the electrical input/outputs of the TIPHON terminals or SCN telephone terminals connected to the TIPHON system. Figure 2 shows in general how this should be done.

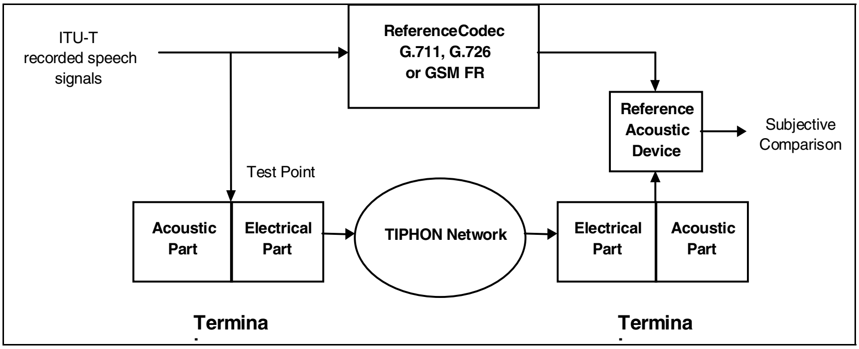


Fig. 2: Methodology for testing TIPHON speech quality

Speech quality shall be measured using the subjective test methodology as defined by ITU-T SG12 until such times as calibrated objective methods are possible. It is planned that these test results will be used in the future to enable predictions of overall performance to be made using a TRM (e.g. the E-Model). It should be noted that the E-model is not a test method.

4 Desktop Videoconferencing Experiments

4.1 State of the Art

4.1.1 State of the Art of QoS in IP Infrastructures

For many years, public network operators regarded ATM as *the* solution for a service integrating broadband network. Conceived as a logical extension of narrowband ISDN, its standardization was influenced by the connection-oriented paradigm, signaling protocols and addressing scheme known from ISDN. While ATM research, development and standardization was concerned with guaranteed QoS for typical broadband ISDN applications, such as video-on-demand and multimedia conferencing as well as with an efficient rate control for data applications (Available Bit Rate), the World Wide Web helped to establish IP networks as *the* carrier for data networking.

The current Internet architecture offers a flexible, but simple connectionless best effort service and this is inadequate for applications sensitive to the QoS provided by the network. For this reason, the IETF has been working on extensions to the current Internet protocol suite in order to enable service guarantees (Resource Reservation Protocol RSVP together with Integrated Services, IIS) or at least differentiation (Differentiated Services, DS).

In Integrated Services, RSVP allows applications to request either Guaranteed or Controlled Load Service for individual flows in an IP network. However, the wide scale deployment of RSVP must be approached with care because the processing of (periodically refreshed) reservation and control messages, the identification of each packet based on the IP header and the handling of per-flow reservation state becomes challenging in backbone routers passed by a huge number of individual flows.

Several ACTS projects (DIANA, SUSIE, BTI, PETERPAN, IthACI and ELISA) have been working on the implementation and optimization of a RSVP over ATM control architecture to integrate IP and ATM while enforcing QoS end-to-end. Their work extends standardized solutions to integrate IP and ATM on a best-effort basis, namely Classical IP over ATM (CLIP), Next Hop Resolution Protocol (NHRP), Multicast Address Resolution Server (MARS), LAN Emulation (LANE) and Multiprotocol over ATM (MPOA) that focus on IP to ATM address resolution to set-up switched best-effort ATM VCs.

The resulting RSVP over ATM architecture is an example of traffic descriptor and QoS parameter based resource reservation that guarantees tight QoS end-to-end. The aforementioned scalability issues are addressed by applying a concept of massive aggregation of flows to a single VC.

In contrast, Differentiated Services architecture achieve scalability by classifying and marking packets by means of the so-called DS field in the IP header once at the ingress to a DS capable IP network. Based on a (relatively static) Service Level Agreement profile negotiated between an Internet provider and an user, traffic will receive a particular per-hop forwarding behavior on routers that interpret the DS field. Without explicit signaling and admission control, DS have the potential to provide relative QoS to so-called behavior aggregates (streams that are marked with the same priority) using simple and scalable mechanisms.

Hence, the Internet is evolving in the direction of a multi-service network that supports several traffic classes, signaling and various other attributes associated with a stream of data. Since various solution will co-exist, QoS applications and services have to build upon a more generic protocol layer, as represented by the H.323 series of recommendations, that is to be translated to native network layers and traffic control capabilities in terminals (with operating systems support) or gateways (network equipment vendors). This approach enables application developers to keep pace with and to make use of the rapidly evolving IP based technologies and network infrastructures.

4.1.2 State of the Art in QoS Evaluation Methodologies, Tools, and Experiments

Today, many of activities address QoS evaluation methods and tools dedicated to broadband networks ; most of them concentrate on objective (quantifiable, measurable) QoS aspects, which are mainly network oriented. Subjective aspects relate to the user’s point of view and are only approached qualitatively.

Objective QoS

The IETF’s work on Quality of Service and performance is primary directed towards developing new protocols which will allow a degree of bandwidth reservation or Quality of Service differentiation. The IP Performance Metrics working group (IPPM) is working on defining a set of standard metrics that can be used to derive a quantitative measure of the quality, performance and reliability of Internet data delivery services.

Measurement and monitoring processes will also take into account work performed in :

- ACTS Project ISABEL: video-conferencing traffic characteristics, ISABEL and Mbone tools,
- Project MEHARI (Spanish local project): techniques and tools for the analysis of Internet services. Currently the MEHARI system is used to do measurements on IP over ATM. New functionality could be as for example LANE/MPOA measurements, end-to-end QoS monitoring, etc.

- Project SABA (Spanish local project): new services and protocols for the Broadband Spanish Academic Network. It is a project on next generation Internet (QoS management, terminal and network related aspects).

Subjective QoS

QoS requirements by the user/customer is the statement of the level of quality of a particular service required or preferred by the user/customer. The level of quality may be expressed by the user/customer in technical or non-technical language.
A typical user/customer is not concerned with how a particular service is provided or with any of the aspects of the network's internal design, but only with the resulting end-to-end service quality. From the user's/customer's point of view, QoS is expressed by parameters that:

- focus on user/customer-perceivable effects, rather than their causes within the network,
- do not depend in their definition on assumptions about the internal design of the network,
- take into account all aspects of the service from the user's/customer's point of view,
- may be assured to a user/customer by the service provider(s),
- are described in network independent terms and create a common language understandable by both the user/customer and the service provider.

Classes of Service

The terminology Class of Service is used to describe a scheme where service types are grouped logically together. This grouping can then be used as the basis for service type prioritization.
Four classes of service are defined by ETSI:

- Class4 : Best quality,
- Class3 : High Quality,
- Class2 : Medium Quality,
- Class1 : Best Effort Quality.

Different (and complementary) approaches are being defined in the two main QoS architectures of the IETF [5]. These architectures will be prevalent in the near future Internet:

- Differentiated Services (DiffServ):
  . Default (DE): best-effort,
  . Assured forwarding (AF): not completed yet,
  . Expedited forwarding (EF): not completed yet.
- Integrated Services (IntServ):
  . Guaranteed: with bandwidth, bounded delay and no-loss guarantees,
  . Controlled load: simulates a best-effort service in a lightly loaded network,
  . Best-effort.

One important issue under work is the mapping of CoS definitions of both standards IntServ and DiffServ.

4.1.3 State of the Art of Multimedia Conferencing Systems

Almost all multimedia or video conferencing systems used today are either based on IP (e.g. CUSeeMe, the MBone tools and Microsoft's Netmeeting) or on ISDN (e.g. Intel ProShare and PictureTel). This leads to a conservative dimensioning where the application on the one hand must be able to deal with data losses and on the other hand must put restrictions on its output bit rate to protect the other services running over the same network from excessive losses.

This scheme works very well for data, but is problematic for real-time audiovisual services. Furthermore the available bit-rate on the Internet is quite low which leads to designs with low resolution (image size) and low frame-rates. With the use of ATM these problems could be reduced or eliminated.

Multimedia conferencing systems using native ATM are not commercial available today, but some research projects have worked in this area. The RACE project R2025 MIMIS have worked on multimedia desktop conferencing for ATM networks, using the early drafts of the ITU-T T.120 series.

4.2 The DIVINE Project

DIVINE is an European consortium regrouping industrial companies, research centers, telecommunications operators, universities and end-users [4]. The DIVINE project was initiated by the European Commission (DGXIII) in the frame of the 4th PC&RD ACTS Program.

The main objectives of the DIVINE project were:

- to be a "market-driven" project through operational field trials involving real end users, to issue meaningful conclusions on the viability of the deployment of multi-point multimedia applications on high speed communication infrastructures,
- to be a "product-oriented" project : the reuse of the results of the project is a key issue. It implies that the DIVINE system has to fit with the user requirements related with functionality, quality of service and price,
- to demonstrate the interoperability between the B-ISDN DIVINE and N-ISDN commercial videoconferencing products,
- to demonstrate on a large scale experimentation the interoperability of European ATM-LAN with the ATM-WAN,
- to promote the standards in broadband desktop videoconferencing applications and to contribute to the standardization work in IMTC, ETSI or ITU-TS [6].

Bandwidth-on-demand should be easily and flexibly controlled by the user. Support of a variety of interface-cards for ATM, graphics, video-input and video compression should be integrated. Access to these peripheral components should be enabled via

hardware interfaces. If such an interface does not exist, at least a standardization should be considered or this should be open to the public.
The DIVINE protocol stacks can be described as follows:

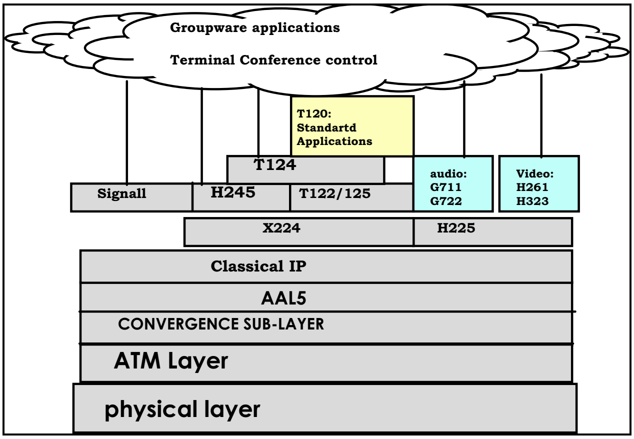


Fig 3: The DIVINE protocol stack

Unicast and multicast videoconferencing is requested, as well as easy interworking with ISDN based videoconferencing facilities.

4.3 Tests and Performance Measurements

The host configuration includes machines, graphic interfaces, audio/video codecs and the network access interfaces includes LAN Ethernet: 10 Mbit/s, 100 Mbit/s, Fast Ethernet, Switched Ethernet, 155 Mbit/s ATM. Quality of Service evaluation relates to end-to-end multimedia applications with LAN-to-LAN interconnection, audio/video conferencing and audio/video transmission and distribution.
The main test functions are the user/network protocol monitoring, the user/network protocol simulation, the traffic load simulation and the errors insertion (bit/cell/frame errors, delays, jitter...).

In user simulation mode, the ATM analyzer acts as any number of user devices which may communicate with a real ATM switch under test. A real ATM switch is attached to the analyzer which may simulate any number of user devices which are called virtual stations. Additionally, the analyzer enables the simulation of any number of user devices which may communicate with the simulated switch or with any other user device. The measurements include:

- estimation of bandwidth allocated to a videoconferencing session,
- videoconferencing behavior under strong network load conditions,
- audio/video quality estimation,
- determination of optimal Quality/Bitrate relations,

- interoperability with ISDN (H320/H323 gateway tests),
- native ATM interworking (ATM access for DIVINE terminal).

Objective measurements were done essentially by observing traffic at the ATM level (Radcom monitoring equipment) and at the Ethernet level (Meterware monitoring equipment), via cable-modems and via ISDN H320 gateway [15].

During our tests, both videoconferencing applications are adjusted for maximum video quality, (the maximum bandwidth allocated for DIVINE is 750 kbit/s). Statistics are extracted with the RADCOM test equipment, connected to the ATM network and the results are depicted for a 1 minute connection (data storage capabilities).

The results can be represented in this following figure:

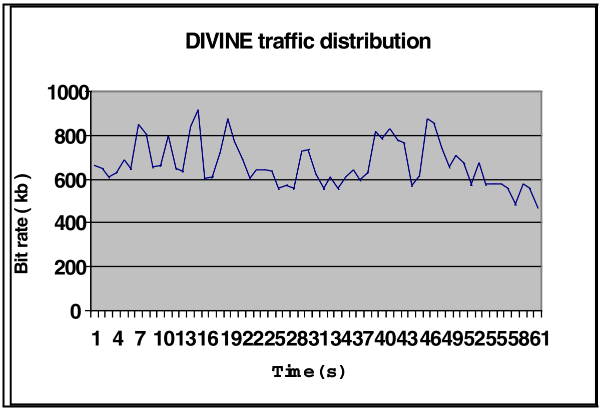


Fig. 4 : DIVINE traffic distribution

Decreasing available bandwidth affects video and audio quality, but the connections are not lost. The maximum quality requests only 450 kbit/s to 900 kbit/s bandwidth allocation and videoconferencing applications request a minimum of available bandwidth to provide acceptable video and audio quality.

Fugitive traffic congestion disturbs the video transmission for short periods of time, but does not affect the videoconferencing session itself. When the available bandwidth increases again, video and audio quality recover their initial values.

5 The Future of VoIP Networks

New operators use more and more VoIP because they can use the same equipment to transmit voice and data over Internet. The development of VoIP implies the integration of the PSTN (Public Switched Telephone Network) network. Therefore gateways that can be used to interconnect SS7 protocol with IP protocol are now available. The three majors protocols for VoIP and SS7 over IP are H.323, MGCP (Media Gateway Control Protocol) and SIP (Session Initiative Protocol).

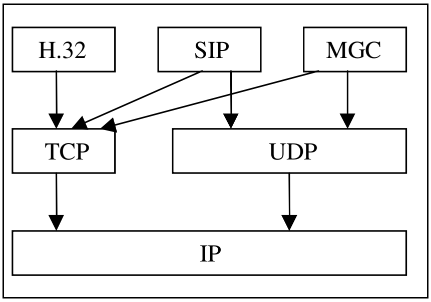


Fig. 5: VoIP and SS7 over IP protocols

The H.323 ITU (International Telecommunication Union) standard is adapted for multimedia conferences. It can be used to transmit voice over ATM, but H.323 cannot really evolve to integrate the SS7 signalization.

The MGCP IETF (Internet Engineering Task Force) standard, which come from the merge of SGCP (Simple Gateway Control Protocol) with IPDC (Internet Protocol Device Control), has been developed to resolve the SS7/VoIP integration. Then, MGCP can be used to offer operational VoIP networks based on PSTN over IP.

The SIP IETF standard has been initially developed for the multimedia communication over the Mbone (Internet Multicast Backbone). SIP offers SS7 over IP mechanisms and can be used instead of H.323. SIP is simpler than MGCP that offer more control mechanisms.

In the future, it will be important to evaluate how SIP or MGCP can be managed by the DiffServ and MPLS protocols.

6 Conclusion

The QoS for VoIP systems and desktop videoconferencing is more and more enhanced. In this article, we present some methods and performance measurements applied to VoIP and to desktop videoconferencing systems. Some methods for the simulations and for the tests are experimented in different projects.

7 References

1. K. Achtmann, K.H. Doring, R. Herber, G. Komp, “An ATM-based demonstration model for multimedia services using different access networks”, Multimedia Applications, Services and Technologies, ECMAST’97, May 21-23, 1997, Milan, Italy, 1-17.
2. G. Armitage, "MPLS: the Magic behind the Myths", IEEE Communications Magazine, Vol. 38, No 1, , January 2000, 124-131.
3. R. Cocca, S. Salsano, M. Listanti, "Internet Integrated Service over ATM: a Solution for Shortcut QoS Virtual Channels, IEEE Communications Magazine, Vol. 37, No 12, , December 1999, 98-104.

4. DIVINE (Deployment of Interpersonal Videoconferencing Systems on IBC Networks) project, ACTS European Project AC035, 1998.
5. G. Eichler, H. Hussmann, G. Mamais, I. Venieris, C. Prehofer, S. Salsano, "Implementing Integrated and Differentiated Services for the Internet with ATM Networks: a practical approach", IEEE Communications Magazine, Vol. 38, No 1, January 2000, 132-141.
6. ETSI/DTR/TIPHON-05001 V1.2.5., "Telecommunications and Internet Protocol Harmonization Over Networks ; general aspects of Quality of Service", 1998
7. B. Goodman, "Internet Telephony and Modem Delay", IEEE Network, Vol. 13, May-June 1999, 8-16.
8. A.M. Grilo, P.M. Carvalho, L.M. Medeiros, M.S. Nunes, "VTOA/VoIP/ISDN Telephony Gateway", 2nd International Conference on ATM, ICATM'99, Colmar, June 21-23 1999, 203-235.
9. ISO/IEC DIS13236 , "Information Technology Quality of Service - Framework," 1996.
10. N. Kroth, L. Mark, J. Tiemann, " A Framework for Testing IP QoS over ATM Networks: Implementation and Practical Experiences", 2nd International Conference on ATM, ICATM'99, Colmar, June 21-23 1999, 203-235.
11. V. Mirchandani, D. Everitt, "Performance of integrated telephony service over standalone LAN and ATM interworked LANs", 22nd Conference on Local Computer Networks, Minneapolis, USA, November, 2-5, 1997, 80-88.
12. D. Newman, "VoIP Gateway: voicing doubts", Data Communications International, Vol. 38, No 12, September 1999, 70-78.
13. Z. Peifang, "Scalability and QoS guarantee in IP networks", 8th International Conference on Computer Communications and Networks, 627-633.
14. X. Scharff, P. Lorenz, "Specification of a Multimedia Application Generator in Telecommunication Systems", 12th International Conference on Computer Applications in Industry and Engineering (CAINE'99), Atlanta, USA, November 4-6 1999, 54-57.
15. H. Tobiet, P. Lorenz, "Performance measurements on an ATM-based Metropolitan Area Network: OASICE Case Study", 1st IEEE International Conference on ATM, ICATM'98, Colmar, June 22-24 1998, France, 410-417.
16. J. Toga, J. Oltt, "ITU-T Standardization Activities for Interactive Multimedia Communications on Packet-based Networks: H.323 and related recommendations", Computer Networks, Vol. 31, No 3, February 199,9 205-223.
17. M. Valino, J. Corchado, "VoIP: the convergence of Networks", Computing and Imformation Systems, Vol. 6, No 3, October 1999, 105-112.
18. P.P. White, "ATM switching and IP Routing Integration: the Next Stage in Internet Evolution", IEEE Communication Magazine, April 1998, 78-83.

Efficient Network Utilization for Multimedia Wireless Networks

Xin Liu, Edwin K.P. Chong, and Ness B. Shroff

School of Electrical and Computer Engineering
Purdue University, Lafayette IN 47907, USA
Tel: +1 765 494-1744, Fax: +1 765 494-3358
{xinliu,echong,shroff}@ecn.purdue.edu

Abstract. In this paper, we present an access scheme to satisfy the QoS requirements for two classes of traffic during the *contention phase* of packet-switched wireless communications. In the proposed scheme, different classes of users contend with other users for resources based on controlled class-dependent permission probabilities. We prove that our algorithm is stable for a large class of arrival processes. Under certain QoS requirements, we derive an upper-bound on the throughput for a general class of random access algorithms. We show that the throughput of our algorithm asymptotically approaches this upper-bound. We also show, through numerical examples, that our algorithm achieves high network utilization.

1 Introduction

The goal of wireless communications is to provide a convenient and economical way for people to transfer all kinds of information, such as voice and data. Compared with circuit switching, packet switching provides more efficient multiplexing of different classes of traffic. In circuit switched networks, when a user is admitted to the network, a certain amount of network resource is assigned to the user and exclusively used by the user until its communication finishes, regardless of whether the user has information to transmit during this period. In packet switched networks, when a new user is admitted, no specific resource is assigned to it. Resources are shared by users in the system. A user only occupies the network resource when it has information to transmit. Consider a phone call as an example. When the user talks, voice packets are generated at a certain rate; when the user is silent, no voice packet is generated. On average, the user talks less than half of the entire call duration. In circuit switched networks, the networks assign the voice user the resource equivalent to its packet rate during talking, hence about half of the resources are wasted. In packet switched networks, when a user does not talk, no resource is assigned to this user; when the user begins talking after a period of silence, the network assigns resource to this user again. Hence, in general, packet switching utilizes network resources more efficiently than circuit switching. Efficiency is very important for wireless networks because wireless bandwidth is scarce. However, wireless packet switching

C.G. Omidyar (Ed.): MWCN 2000, LNCS 1818, pp. 108–122, 2000.

suffers from access problems in the uplink. In other words, when a user becomes active, it has packets to transmit and no network resource is assigned to it, the user has to compete with other users to gain the access to network resources. To solve this problem, a variety of contention and reservation medium access control (MAC) protocols have been widely used in the area of communication networks [2, 3, 4, 5]. Typically, there are two transmission phases:

1. Newly activated users compete to gain access to the networks. The first packet of a newly activated user is transmitted through the network using some random access protocols; i.e., contention-based communications. This first packet may be a packet in a special form or a normal data packet. In this paper, we call the first packet a *request*. If the first packet is lost during transmission, or is received in error, then it is retransmitted until successful.
2. Following the first successful contention-based transmission, subsequent transmissions are scheduled contention-free using a scheduling strategy.

We call the first phase the *contention phase* and the second phase the scheduling phase. In this paper, we focus on the contention phase of communications. In packet switched wireless networks, the contention phase may exist throughout the whole communication period, and not only during the admission period. Every time a user becomes active (say, a user begins talking after being silent), at that very moment, because no resource is assigned to the user, the user has to inform the base station about its resource requirement through contention-based communication. Hence, contention-based communication plays an important role in packet-switched wireless networks.

In packet switched networks, admission control and resource allocation are used to provide QoS. In general, admission control is based on the resource allocation scheme. In wired networks, resource allocation is implemented by smart scheduling schemes. However, smart scheduling is not enough to provide QoS for wireless networks, where contention plays an important part. For example, we want to provide delay guarantee to real-time traffic in wireless networks. When a user begins talking, it first sends its request to the base station through random access; i.e., contention-based transmission. Then the base station schedules the traffic after it receives a resource request from the user. Therefore, the user experiences delay caused by contention plus the delay caused by scheduling. To guarantee the delay experienced by the user, we need to guarantee the delay in both contention phase and scheduling phase. During the scheduling phase (if one actually exists in the given implementation), smart scheduling strategies can be used to provide delay guarantees. However, we also need algorithms in the contention phase to provide delay guarantees to users. To provide QoS in the contention phase is intrinsically difficult due to the nature of random access. While there is a significant body of work on the development of effective scheduling and admission control policies to ensure QoS, there is very little work done in implementing QoS during the contention phase of communication.

In this paper, we present an algorithm that implements QoS requirements for two classes of traffic in the contention phase of packet switched time-slotted wireless networks. Controlled time-slotted ALOHA is the random access algorithm considered in this paper. Two traffic classes, voice and data, are considered. We consider only two classes for simplicity of exposition, convenience of calculation and explanations, although more classes can be similarly considered. We assume that voice users have *delay requirements* and data users do not have such requirements.

In wire-line networks, if two or more users transmit at the same time through the same media, usually all of the transmissions are assumed to have failed. However, this assumption may be unnecessarily pessimistic in the mobile radio environment, where the received packets at the base station are subject to the near/far effect and channel fading. Packets from different users in the same slot may arrive at the base station with different power levels and the base station may successfully decode one or more packet. This is referred to as *capture.* Due to the page-limit requirement of this conference, we only present the QoS algorithm for systems that do not exploit capture. However, the proposed algorithm works for systems with capture too [6]. It is obvious that the system throughput will be improved if the system explores capture. However, unfairness exists between near and far users due to the nature of radio transmission. To achieve fairness and good throughput, we present a distance-dependent permission probability scheme that require users at different distances from the base station to transmit with different probabilities to provide certain delay guarantees, *distance fairness*, and good throughput. In summary, if we do not consider the ability of capture, the QoS requirement is presented in terms of delay. When we consider capture, the QoS requirement is explained in terms of delay and distance fairness.

This paper is organized as follows. In Section 2, we describe the system model. We present and analyze the QoS algorithm in Section 3. An upper-bound for the throughput is derived, under certain QoS requirements, for a general class of random access algorithms. The throughput of our algorithm asymptotically approaches this upper-bound. Simulation results are provided in Section 4. Conclusion and future work are presented in Section 5.

2 System Model

In this section we describe the system model. There is a base station with mobile users in its coverage area. We consider the uplink of a time-slotted system and focus on the contention phase of communication. We assume that time is divided into frames and each frame consists of request slots. Each request slot is large enough to contain a fixed size request. The base station monitors and controls the contention phase in the system. *In the following, when we mention users we mean newly activated users with requests to transmit, except otherwise specified.*

At the beginning of a frame, the base station broadcasts a permission probability for each class of users through a non-collision error-free signaling channel. A user decides whether or not to transmit in a request slot in the frame according to the permission probability of its class broadcasted by the base station. Different classes of users may have different permission probabilities.

We assume that a user can transmit at most once in a frame. There are request slots in each frame. The parameter, , determines how often the base station updates its control parameters, and how long a user waits before it retransmits. In practice, the larger the value of , the less the signaling, the better the estimation of the number of users, however, the longer the delay.

In some cases, we prefer a large value of . An example of such a scenario is in satellite communications. After the contention of a time slot, a user cannot know immediately whether its request is successfully received by the hub station. In satellite communications, the round trip delay is relatively large. For instance, the propagation delay is around 20–25ms for LEO (low earth orbit) systems [8]. An immediate ack from the the hub is impossible. Furthermore, the coverage area of satellite communications is relatively large, it is difficult for an earth station to detect whether its transmission is successful. Hence, a large value of may be suitable for such a case. In other cases, a small value of could be favored. A good example of such a case is a local wireless network, where the sum of the round trip delay, and processing time, etc., is small. A user transmits, then waits for acknowledgment. If the user does not receive an acknowledgment from the base station in the predetermined waiting time, it assumes that the transmission has failed. The user could retransmit it in the next frame. The extreme case is when $= 1$; i.e., a user can retransmit its request in the next request slot. In the extreme case $= 1$, the scheme studied in this paper becomes the pure priority scheme; i.e., when there are voice users, no data user transmits, and when there is no voice user, data users transmit. However, even in a wireless LAN, it is not necessary to adopt such a small value of . Usually, the requests are much shorter than normal data packets. Hence, the delay caused by several request slots are tolerable in order to reduce the cost of extensive signaling.

In this paper, we assume that the system is not capable of correctly deciphering any transmissions when two or more overlapping transmissions arrive in the same slot; i.e., if two or more users transmit their requests through the same request slot in a frame, neither of them can be successfully received. This situation is called collision.

We assume that a request is never discarded; i.e., a user always retransmits its request until it is acknowledged by the base station that its request has been received successfully. While the request of a user is delayed, some packets may be buffered at the user. In real-time applications, human factors may decide whether to send a delayed packet or to drop it. This issue is irrelevant in our scheme. Furthermore, we assume that the acknowledgment is error-free and the base station uses a scheduling strategy to decide when the active user should transmit in the reservation phase of communication.

3 The QoS Algorithm

We first present the QoS algorithm with restriction to the delay requirement of voice users. We, then, analyze the throughput and stable condition. Finally, we derive a throughput upper-bound under the QoS requirement for a large class of random access algorithms.

3.1 Algorithm

Let $_v$ ($_d$) denote the permission probability that a voice (data) user transmits in a request slot in a frame. In this paper, the permission probabilities, $_v$ and $_d$, are used to stabilize the ALOHA system, to achieve good throughput, and to provide QoS guarantees. The use of permission probabilities to stabilize ALOHA is not a new idea. Permission probabilities are also used to provide priority to voice users in [4, 7]. In the literature, there are algorithms, centralized and decentralized, to estimate the number of users in the system. All these algorithms can be used in our scheme. Hence, we focus on how to use the permission probabilities to satisfy QoS instead of how to estimate the number of users. During the analysis we assume that the base station knows the precise numbers of voice users and data users in each frame. Knowing this information is the ideal condition of the algorithm. Practically, we use a Kalman filter to estimate the numbers of voice users and data users with requests in each frame. We show through simulations that using a Kalman filter for the estimation provides very good results.

As mentioned before, a user can transmit at most once in a frame. We do not distinguish between newly arrived and retransmitted users. The base station broadcasts $_v$ and $_d$ at the beginning of frame . A voice user randomly selects a request slot to transmit in this frame with probability $_v$, as would a data user with probability $_d$. All users select and transmit independently. The base station acknowledges those users whose requests have been successfully accepted at the end of frame . Users that have not been acknowledged assume that their requests have not been successfully transmitted. They retransmit in the next frame. The base station estimates the number of users in the system, calculates $_v$ and $_d$ for frame $+ 1$, and so on. It is easy to prove that the throughput is maximized when users transmit in each frame [6]. However, this throughput may come at the cost of excessive delay for voice users. Hence, we need to develop a scheme that attempts to maximize throughput subject to a given level of delay requirement for voice users.

A good measure of QoS is the delay experienced by a user before its request is successfully received by the base station. However, the precise delay distribution of voice users is very difficult to find in this context. Thus, we define an average success probability, $\bar{P}_s$, as the QoS measure used in this paper. Suppose the system has reached steady state. When a voice user becomes active, on average, it transmits its request successfully with probability $\bar{P}_s$, given by

$$\bar{P}_s := \quad [\ {}_s(\quad {}_v \quad {}_d)] = \quad \underset{i,j}{\ } \quad {}_s(\quad j)\ (\quad j) \tag{1}$$

where $_s($ $j)$ is the probability that a voice user transmits its request successfully in a frame in steady state when there are voice users and j data users in the system, and $(\ j)$ is the steady state distribution that voice users and j data users are in the system.

Our QoS requirement for voice users is $\bar{P}_s - \ _0$, where $_0$ is the given delay threshold. Roughly speaking, the contention delay of a voice user is geometrically distributed with parameter $\bar{P}_s$; i.e., the distribution of access delay is approximated by $P(\ = \ ) = \bar{P}_s(1 - \bar{P}_s)^{x-1}$. When the correlation of the numbers of users between cells is small, the approximation is good. If the number of users arrived in each frame is independent, then the larger the , the better the approximation. In Section 4, we show that the distribution of voice users from simulations is well approximated by a geometric distribution (see Figure 1).

The QoS algorithm is described as follows. Suppose that the base station knows that $_v$ voice users and $_d$ data users are in the system. Then, the permission probabilities of voice users and data users are

$$_v = \min(1 \ \frac{\ }{\ _v})$$

$$_d = \begin{array}{ll} \min \ 1 \ \frac{(C-N_v)^+}{N_d} & : \ \text{if} \ \ _v \ \ 0 \\ \min \ 1 \ \frac{M}{N_d} & : \ \text{if} \ \ _v = 0, \end{array} \tag{2}$$

where

$$(\ )^+ = \begin{cases} & : \ \text{if} \ \ - 0 \\ 0 & : \ \text{otherwise.} \end{cases}$$

Note that is a tuning parameter used to satisfy the QoS requirements of voice users. So the algorithm does the following. If the number of voice users in the system is less than , all voice users can transmit freely. In this case, data users may or may not be allowed to transmit. If the number of voice users in the system is greater than , then a voice user is allowed to transmit based on the outcome of the toss of a biased coin with probability $_v$ of success. In this case, no data users are allowed to transmit. Before we illustrate how to calculate , we first make a few observations:

- Data users yield to voice users the right to access request slots.
- The parameter satisfies $0 - \ - \ $. The expected number of data users to transmit is $(\ - \ _v)^+$. The total throughput is maximized when $\ = \ $. The larger the value of , the higher the throughput, and the larger the delay of voice users. Hence, there is a tradeoff between the throughput of the system and the delay requirement of voice users. When the QoS requirement is stringent, is small, data users are allowed to access request slots with lower probability, and voice users have a higher probability to succeed in a frame.
- When there is no voice user; i.e., $_v = 0$, the value of $_d$ is set to maximize the throughput.

The tuning parameter can be calculated theoretically. A two dimensional Markov chain is used to calculate the steady-state distribution. Suppose that we know the distribution of the arrival process. Let $=\ ,0 - \ - \ $. Transmission probabilities between states are determined by (2) and the arrival process. Hence, $(\quad)$ can be calculated and so can $\bar{P}_s$. Since $\bar{P}_s$ is a monotone decreasing function of , denoted as $\bar{P}_s(\ )$ for $0 - \ - \ $, the parameter is the unique root of $\bar{P}_s(\ ) = {}_0$, which can be obtained easily using standard zero-finding algorithms. If $\bar{P}_s(0) \quad {}_0$, the QoS requirement cannot be satisfied. In other words, even without data users, the delay caused by the contention among voice users are still larger than required when $\bar{P}_s(0) \quad {}_0$.

Practically, there is a very simple approximation for . Let ${}_0$ satisfy

$$1 - \frac{1}{\quad}^{K_0 - 1} = {}_0 \tag{3}$$

If ${}_0$ is not too small compared to and the fraction of voice users is not too large, then ${}_0$ is a good approximation of . In this case, the number of voice users in the system in steady state is seldom larger than ${}_0$. Therefore, the average delay $\bar{P}_s$ is:

$$\begin{aligned}\bar{P}_s &= \ (\ _s) = \ (\ _s(\ ) - - \ )\ (\ - \ ) + \ (\ _s(\ ) - \ )\ (\quad) \\ &\approx (1 - \frac{1}{\quad})^{C-1} = (1 - \frac{1}{\quad})^{K_0 - 1} = {}_0\end{aligned}$$

In fact, if ${}_0 - 0\ 5$ and the fraction of voice users is less than 70%, $\approx {}_0$ is a good approximation. We set $= {}_0$ in simulations in Section 4 and find that it works well.

We, next, analyze the algorithm. First, we calculate the throughput. Second, we prove that the algorithm is stable for a large class of arrival processes. Then, we derive an upper bound on the throughput of random access algorithms under the QoS requirement $\bar{P}_s - {}_0$. We show that the throughput of our algorithm asymptotically approaches the upper-bound.

3.2 Throughput

Suppose that there are users transmitting in a frame. Each user selects one of the request slots randomly and independently. When only one user transmits in a request slot, we assume that the transmission is successful. When two or more users transmit in the same request slot, we assume that neither of the transmission is successful. The throughput, ${}_k$, is defined as the average number of requests that are successfully transmitted in a frame and ${}_k$ is the probability that a user transmits successfully. We then have

$${}_k = \quad 1 - \frac{1}{\quad}^{k-1}$$

$${}_k = \frac{k}{\quad} = \quad 1 - \frac{1}{\quad}^{k-1}$$

We consider the throughput under three conditions:

1. When $_v -$, each voice user transmits in a request slot with probability $_v = \min(1 \quad _v)$ and no data user transmits. The throughput is:

$$(\quad_v \quad_d) = \overset{N_v}{\underset{i=0}{\;}} \quad {}_iP(\text{ voice users transmit in this frame})$$

$$= \quad_v \;_v \;\; 1 - \frac{_v}{\;}^{\;N_v-1} \tag{4}$$

2. When $_v$, each voice user transmits in a request slot with probability 1 and each data users transmits with probability $_d = (\quad - \quad _v) \quad _d$. Therefore,

$$_s(\quad_v \quad_d) = \overset{N_d}{\underset{i=0}{\;}} \quad {}_{i+N_v}P(\text{ data users transmit in this frame})$$

$$= \;\; 1 - \frac{1}{\;}^{\;N_v-1} \quad 1 - \frac{_d}{\;}^{\;N_d}$$

The throughput consists of successfully transmitted voice and data requests:

$$(\quad_v \quad_d) = \overset{N_d}{\underset{i=0}{\;}} \quad {}_{i+N_v}P(\text{ data users transmit in this frame})$$

$$= \;_v \;\; 1 - \frac{1}{\;}^{\;N_v-1} \quad 1 - \frac{_d}{\;}^{\;N_d}$$

$$+(\quad - \quad _v) \;\; 1 - \frac{1}{\;}^{\;N_v} \quad 1 - \frac{_d}{\;}^{\;N_d-1} \tag{5}$$

3. When $_v = 0$, data users transmit with probability $_d$, $_d = m \quad (1 \quad\quad _d)$, to maximize the throughput.

$$(0 \quad _d) = \;_d \;_d \;\; 1 - \frac{_d}{\;}^{\;N_d-1} \tag{6}$$

3.3 Stability Analysis

We now prove that our algorithm is stable with a fairly weak assumption on the arrival process. We consider a system with a unique stationary distribution as a stable system. We use Pake's Lemma to find a sufficient condition for the system to be stable [9].

Lemma 1 (Pake's Lemma). *Let* $-\;_k \quad = 0 \;\; 1 \;\; 2$ —— *be an irreducible, aperiodic homogeneous Markov chain with state space* $-0 \;\; 1 \;\; 2$ ——. *The following two conditions are sufficient for the Markov chain to be ergodic.*

a) $-(\quad_{k+1} - \quad_k - \;_k = \quad) - \quad - \quad -$

b) $\limsup\limits_{i} \quad (\quad_{k+1} - \quad_k - \;_k = \quad) \quad 0$

Note that an irreducible, aperiodic, ergodic Markov chain has a unique stationary distribution.

Let $_k$ be the total number of users that arrive in the th frame. Suppose that $-{}_k \quad = 0\ 1\ 2$ —— are random variables with mean value . Let $_k$ be the number of users (voice users and data users) at the beginning of the th frame, then ${}_k = {}_v + {}_d$. Let $({}_k)$ be the number of users whose requests are successfully transmitted in the th frame. We now prove that $-{}_k \quad =$ $0\ 1\ 2$ ——is ergodic using Pake's lemma. We have

$${}_{k+1} = {}_k + {}_k - ({}_k)$$

So, for any ,

$$\begin{aligned} &-(\ {}_{k+1} - {}_k - {}_k = \ )-= -(\ {}_k - (\ {}_k) - {}_k = \ )- \\ &= -(\ ) - [\ (\ )]— -(\ )-+-[\ (\ )]— \quad + \quad - \end{aligned}$$

Hence, condition (a) of Pake's lemma is satisfied.

To satisfy condition (b) of Pake's lemma, we require that

$$\begin{aligned} &\limsup_i \ (\ {}_{k+1} - {}_k - {}_k = \ ) \\ &= \limsup_i \ (\ {}_k - (\ {}_k) - {}_k = \ ) \\ &= \limsup_i (\ - \ [\ (\ )]) \quad 0 \end{aligned}$$

So

$$-\liminf_i \ [\ (\ )] \tag{7}$$

is a sufficient condition for the system to be stable.

In our QoS algorithm, when there are $_v$ voice users and $_d$ data users, the total number of users is $= {}_v + {}_d$. Then, there exists an such that for all $-$, we have [6]:

$$(\ {}_v \ {}_d) - \quad 1 - —— ^{i-1}$$

Hence,

$$(\ ) - \quad 1 - —— ^{i-1}$$

Then

$$\liminf_i \ [\ (\ )] - \liminf_i \quad 1 - —— ^{i-1} = \ ^{-\frac{C}{M}} \tag{8}$$

Hence, from (8), $- \quad ^{-C/M}$ is the sufficient condition for the system to be stable under the QoS requirement $\bar{P}_s - {}_0$, where is the arrival rate. Note that in the special case $=$; i.e., the system is designed to achieve the maximum achievable throughput, (8) becomes:

$$\liminf_{i} \ [\ (\ )] = \liminf_{i} \quad 1 - \frac{1}{\ }^{\ i-1} \quad = \quad {}^{-1} \tag{9}$$

The sufficient stable condition is $\quad {}^{-1}$, which is exactly the stable condition for slotted ALOHA. Furthermore, there is no bistable point in the system because the throughput does not decrease when the number of blocked users in the system increases.

3.4 Upper Bound on Throughput

We consider the QoS requirement as $\bar{P}_s - \ _0$. With this restriction, we derive an upper-bound on the throughput for random access algorithms satisfying the following two assumptions. First, all users transmit in request slots randomly and independently. Second, each user transmits in at most one request slot in each frame. Let Ω be the set of all such random access algorithms.

We consider the throughput under two conditions. Condition 1: there is at least one voice user in the system. Condition 2: there is no voice user in the system. First, we consider the throughput under Condition 1. Let denote the total number of users that transmit in this frame, $1 - \quad - -$. The probability that the voice user successfully transmits its request in this frame is .

$$= \begin{cases} \ _X & : \text{ if the user transmits in this frame,} \\ 0 & : \text{ otherwise,} \end{cases}$$

where

$$\ _X := \quad 1 - \frac{1}{\ }^{\ (X-1)}$$

Note that

$$(\ _X) = \qquad 1 - \frac{1}{\ }^{\ (X-1)} \qquad - \quad (\ ) = \bar{P}_s - \ _0 \tag{10}$$

Let $\ _1$ be the throughput given that there is at least one voice user in the system. Then,

$$\ _1 = \qquad 1 - \frac{1}{\ }^{\ (X-1)} \tag{11}$$

We want to maximize (11) with the constraint (10). Let $\quad = (1 - 1 \quad)^{(X-1)}$ So

$$(\ ) - \bar{P}_s = \ _0 = \ 1 - \frac{1}{\ }^{\ K_0 - 1}$$

Let $(\ ) = - \frac{\ln y}{\ln\left(1-\frac{1}{M}\right)} + 1$ which is a strictly convex function. By Jensen's inequality [1],

$$_1 = (-\ (\ )) - -\ (\ (\ )) = {}_0\ 1 - \frac{1}{\ }^{K_0-1} =: {}_C \qquad (12)$$

Next, we consider the condition 2; i.e., no voice user is in the system. Let ${}_0^{\alpha}$ be the throughput of a random access algorithm when there is no voice user in the system. Let ${}_0^{m} = \max - {}_0^{\alpha} \quad - \Omega -$. Let ${}_{\alpha}$ denote the probability that no voice user is in the system of a random access algorithm . Let $P_0 = \max - {}_{\alpha} \quad - \Omega -$. For algorithm , let P_1^{α} be the probability that there is at least one voice user in the system. Hence, $1 - P_1^{\alpha} - P_0$. The throughput of algorithm is given by:

$$\begin{aligned} &= {}_1P_1^{\alpha} + {}_0^{\alpha}(1-P_1^{\alpha}) - {}_CP_1^{\alpha} + {}_0^{m}(1-P_1^{\alpha}) \\ &= {}_C + (1-P_1^{\alpha})({}_0^{m} - {}_C) - {}_C + P_0({}_0^{m} - {}_C) =: {}_{max} \end{aligned} \qquad (13)$$

Therefore, ${}_{max}$ is the upper-bound on the throughput of random access algorithms in Ω (algorithms such that all users transmit for request slots randomly and independently, and each user transmits for at most one time slot in a frame). This upper-bound is not restricted to the $(\ {}_v \quad {}_d)$ strategy used in this paper.

The above upper-bound, ${}_{max}$, may not be tight. We compare ${}_1$ with ${}_C$. Since is a strictly convex function, (12) achieves equality when $= (\ )$ with probability 1. Hence, the upper-bound ${}_C$ is only achievable if $=$ with probability 1; i.e., there are always exactly users transmitting in each frame. However, may not be an integer and it may not be possible to let exactly users transmit in random access algorithms. So ${}_{max}$ may not a tight upper-bound. We try to approach the upper-bound by assigning ${}_v$ and ${}_d$ such that $(\ {}_v\ {}_v + {}_d\ {}_d) =$ in our scheme.

Next, we show that

$$\lim_{M} \frac{(\ {}_v \quad {}_d)}{{}_{max}} = 1$$

when there are enough users in the system.

With some tedious algebra [6], we can show that

$$(\ {}_v \quad {}_d) - {}_C\ 1 - \frac{1}{\ }$$

when

$$_v + {}_d -$$

Recall that P_0 is the maximum probability that there is no data user in the system. Let ${}_0$ be the probability that there is no new voice user with a request in a frame. Then, $P_0 = {}_0P$(all voice users with requests transmit successfully by the end of a frame in steady state). In practice, P_0 is small when is large; i.e., in a large frame, it is unlikely there is no voice user in the frame. For example, if the arrival process of voice users is a Poisson process with mean , then $P_0 - \ _0 = \ ^{-vM}$. Suppose $P_0 - 0$ as $- -$. We have

$$\frac{(\ _v \ \ _d)}{_{max}} - \frac{_C\,(1 - \frac{1}{M}(}{_C + P_0(\ _0^m - \ _C)} - 1$$

Hence, the throughput of the presented QoS algorithm asymptotically approaches the upper-bound. In other words, when there is at least one voice user in the system, the throughput of our QoS algorithm approaches $_C$. Furthermore, as goes large, the probability that there is no voice user in the system goes to zero. So the throughput of our QoS algorithm asymptotically approaches the upper-bound.

4 Simulation Results

In this section, we provide simulation results of the proposed scheme. For all simulations in this section, we set $= 20$; i.e., there are 20 request slots in each frame. For each figure, we run simulation for 100 000 frames in a single-cell. We assume the arrival processes of voice users and data users are independent Poisson processes with the same average rate.

At the beginning of each frame, the base station announces $_v$ and $_d$, the numbers of voice users and data users in the system. (The announced numbers are estimated by the base station in the practical approach.) Knowing $_v$ and $_d$, each user decides its transmission probability according to (2). With this probability, the user selects and transmits in a request slot in the frame. If the user is the only one transmitting in its request slot, its transmission is successful. Otherwise, the user has to wait for the next frame to retransmit and its delay is increased by one. The unit of delay is frame.

Figure 1 indicates the delay distribution of a voice user when $_0 = 0\ 6$, where $_0$ is the required success probability. We can see that the delay distribution of a voice user is well approximated by a geometric distribution when the numbers of new arrived users at different frames are independent and $= 20$. Hence, in other figures, we use the success probability as the delay performance measure.

Figures 2 and 3 illustrate the performance of the proposed QoS algorithm. Figure 2 indicates the delay performance of voice users. The delay performance is shown by the average probability of success. Simulations are run under both the ideal condition and the practical condition. By the ideal condition, we mean that the base station knows the exact numbers of voice and data users in the system. In practice, a Kalman filter is used to estimate the numbers of users. The Kalman filter approach is implemented with two threshold values. We use (3) to approximate . In the ideal condition, (3) offers a pretty good approximation. With $= \ _0$ in the Kalman filter approach, $\bar{P}_s$ is less than the QoS requirement due to estimation errors. Thus, in practice, we should use a smaller threshold value than the one calculated under the ideal condition, which is represented by the curve with $= 0\ 9\ \ _0$. Figure 3 shows the throughput performance. It is

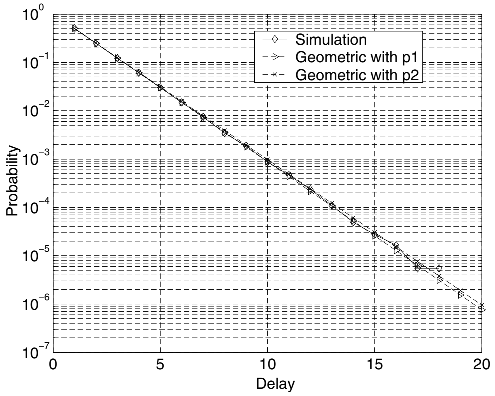


Fig. 1. Delay distribution of a voice user when = 20, 1 is the reciprocal of the average delay of voice users, and 2 is the average probability of success of voice users.

obvious that the throughput decreases with the increase of $_0$. We compare the throughput in the ideal condition with the practical approaches. As expected, the Kalman filter approach with the smaller has less throughput, illustrating the tradeoff between the throughput and QoS. We use the probability of no new voice user in a frame, $_0 = {}^{-r\lambda}$, as the upper-bound of P_0, where P_0 is the probability of no voice user in a frame. Hence, $_0$ is used to calculate the upper-bound of throughput shown in Figure 3, which results a looser upper-bound than that in (13). However, we still note that in most cases, the throughput in the ideal condition is quiet close to the upper-bound in the figure.

5 Conclusions

We present a random access scheme that provides certain QoS guarantees during the contention phase of communication. Permission probabilities are used to provide QoS for two traffic classes, voice users and data users. The same idea can be extended to multi-class users. The QoS requirement of voice users is defined as $\bar{P}_s$, the average success probability of voice users. For a predetermined QoS measure $\bar{P}_s$, a threshold is calculated such that a voice users has an average success probability larger or equal to $\bar{P}_s$. We prove that the algorithm is stable with a weak assumption. We derive the upper-bound of a general class of random access algorithms under the QoS requirement in term of $\bar{P}_s$ and show that the studied algorithm asymptotically approaches the upper-bound. The analysis is based on the QoS algorithm without capture. Note that the QoS algorithms with and without capture are the same in essence except that the success probability is higher when capture is considered [6].

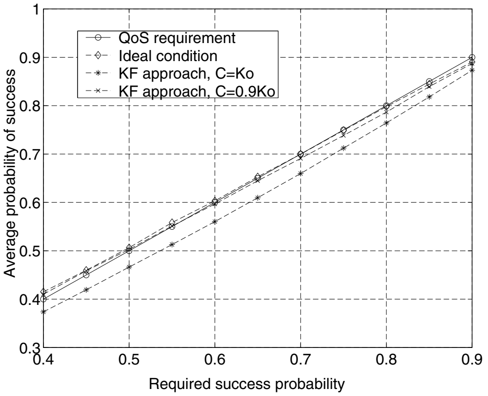


Fig. 2. Delay performance without capture for = 20 with 50% voice users. In the legend, KF denotes Kalman filter.

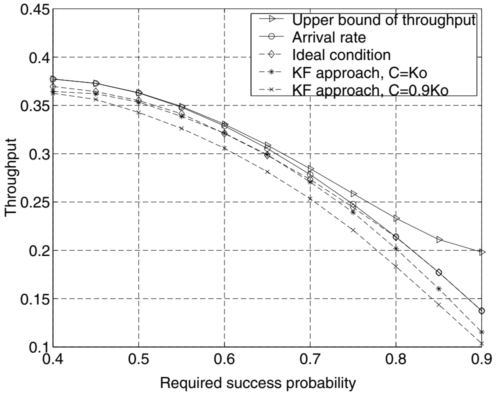


Fig. 3. Throughput without capture for = 20 with 50% voice users. In the legend, KF denotes Kalman filter.

In wireless networks, providing QoS during contention phase is important to support bursty traffic. It is quite different from the wire-line scenario. So existing methods such as using in ATM do not apply directly. There would be large research space for this topic.

References

[1] P. Billingsley, *Probability and Measure.* Wiley, 1985.
[2] C. Bisdikian, “A review of random access algorithms,” *IBM Res. Rep.*, no. RC20348, Jan. 1996.

[3] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard, and B. Ramamurthi, "Packet reservation multiple access for local wireless communications," *IEEE Trans. Commun.*, vol. 37, no. 8, pp. 885–890, 1989.

[4] W. S. Jeon, D. G. Jeong, and C.-H. Choi, "An integrated service MAC protocol for local wireless communications," *IEEE Trans. Veh. Technol.*, vol. 47, no. 1, pp. 352–363, 1998.

[5] R. LaMaire, A. Krishna, and H. Ahmadi, "Analysis of a wireless MAC protocol with client-server traffic and capture," *IEEE J. Sel. Areas Commun.*, vol. 12, no. 8, pp. 1299–1313, 1994.

[6] X. Liu, E. Chong, and N. Shroff, "An access scheme to provide qos in packet-Switched wireless networks," Tech. Rep., Purdue University, 2000.

[7] K. Mori and K. Ogura, "An adaptive permission probability control method for integrated voice/data CDMA packet communications," *IEICE Trans. Fundamentals*, vol. E81-A, no. 7, pp. 1339–1348, 1998.

[8] H. Peyravi, "Medium access control protocols performance in satellite communications," *IEEE Communications Magazine*, vol. 37, no. 3, pp. 62–71, 1999.

[9] R. Rom and M. Sidi, *Multiple access protocols : performance and analysis.* New York : Springer-Verlag, 1990.

Mobility Management and Roaming with Mobile Agents

o n h nh[1] v n n[2] n n . u [3]

[1] ri sson orw y ppli s r n r . ox 1 1 illin s
orw y l + 7 1 00
etodvt@eto.ericsson.se
[2] Univ rsi y o slo Uni . ox 70 007 j ll r
orw y l + 7 1 70
sverrest@ifi.uio.no
[3] l nor . ox 701 01 0 slo orw y l + 7 77 99
Jan.Audestad@telenor.no

Abstract. n is p p r w propos mo il n on p n us in si n n impl m n ion o i l us r mo ili y. p p r s r s wi is ussion o r i ion l mo ili y on p s; rmin l n p rson l mo ili y. on p o i l us r mo ili y is in ro u w i n s n s n x nsion o s wo on p s r sui or n s o mo il us r. n un m n l m nisms or mo ili y suppor r s ri n ow y n impl m n y mo il n s. r i lly mo il n s w i n om in ll s m nisms in flxi l w y is n is uss . p p r on lu s wi summ ry o w y mo il n s r w ll sui o provi mo ili y sup por n som su s ions or u ur s u i s.

Background:
The work described in this paper is a result from the mobility part of the TELEcom REsearch Program TELEREP, carried out at the Ericsson Norway Applied Research Center (NorARC) in collaboration with the University of Oslo, Center of Technology at Kjeller Unik [1]. The program has been established to obtain practical experience with Distributed Object Computing methods such as ODP/TINA [2], [3], [4], CORBA[5], mobile code in the implementation of TMN and IN functionality, and to study how mobility and security can be provided in the DPE environment.

1 Introduction

h h g ow h o h ompu n u y ompu how ng up v ywh . o y mo p opl n w h num o ff n ompu ly h o hom o wh n v ll ng. ll h ompu u know h k ng h mo l ompu on h o o ng mo ompu om h ng ompl ly ff n h n o ng wo k om h no m l nv onm n . n o h

C.G. Omidyar (Ed.): MWCN 2000, LNCS 1818, pp. 123–137, 2000.

wo h u on o h mo l u om ng l. o l m l k m ng ppl on un m l u n n un v l l v nom l. u h ng ng m n l n n wo k po n mo o n ng u ng n m on um ng k o v h no m lly u .

h h n mu h op m m u oun ng h on p o mo l g n . Mol g n l k h u mov oun n h og n ou n wo k . h *mobile agent* w m n ompu p og m h u onomou ly on h l o pon o o g n z on n no oun o h y m wh g n x u on n y h M n .

h n n on o h p p o om n h u ' n o mo l y w h mo l g n hnology w h h g p om n u l n mo l y on p h h no y n l z n l g omm l ppl on . n o o o o w w ll x m n g n l mo l y on p o h un m n l mol y m h n m u o mpl m n h m h n m p n . h n h on p o g lly mo l g n n how h o h mo l y m h n m n uppo y g n u .

2 Mobility Concepts

h on w ll p n h on p o m n l n p on l mo l y wh h n h v n p n n ly o h hnology u o h mpl m n on. ow v mpl m n ng h mo l y on p no nough o p ov u n ly mo l y v . h o p opo h on p o l u mo ly o x n m n l n p on l mo l y.

Terminal Mobility

m n l mo l y h l y o p ov v o m n l h mov oun o onn o v y o n wo k po n (. o l o h n wo k h m n l h o onn o . o p ov n wo k h mu x n o on w n h m n l n h . h o on h o yn m o llow o m n l mo l y. h h phy l h h h k n o m n l h n u h n h yp o v h n p ov . n l ommun on m n l mo l y h n uppo y k p ng h u ' u n onnuou ly up lo h u ' hom om n h o Mo l n M. h llow o m n l mo l y u h v h h u n h v n v y l m . h ompu ng n u y h h mo ou l uppo ng m n l mo l y. On l g p h m n g m n o h n l o h u h m lv wh h l m h u uln o m n l mo l y.

Personal Mobility

- n p on l mo l y h l y o l h u v om ny m n l ny lo on . k h n on l l u h n h o m n l on p on l l ommun on n n h n wo k mu p ov v o ng o h u ’ v p o l . w ll u h n on n h p p . o l o p ov v o u onn o m n l on n wo k h mu x no on w n h u n h m n l. h o on mu yn m n o o n u p on l mo l y. u m gh h v p o l . h p o l o h u ’ p n n h mu x om omp l y w n h p o l n h p l o h m n l. o n n h u qu v o on u h m n l n only off vo p l o v om omp om mu m o h m n l mu h qu o v . h p on l mo l y lly n u h h u n o v o h n wo k om num o ff n m n l onn o v y o . ow v h v no nough o mo u . h y l o n o h v o h u om z ppl on n p o l . h m p on l mo l y o no n u h n up o h u h mlv o xp n h v off .

Ideal User Mobility

v n wh n m n l n p on l mo l y om n h mo l u m gh ll no . p opo h on p o l u mo l y wh h n o poh on p o m n l n p on l mo l y wh l ng n p n y on op o h on p n x n ng h num o ppl on n p ol h n . lly h ff n ompu ng nv onm n houl p o h u n o h u h v ng o p o h m. *Ideal user mobility* n h n n h l y o p ov h m ppl on n p o l n p n ly o h u nywh ny m n w h h m look n l. h on p o l u mo l y u n h p p h om n log o h M ’ u l om nv onm n (on p . n *system concept for personalized service portability across network boundaries and between terminals* 9 10 . ow v h on p o u only on v wh l ou on p o l u mo l y omp o v ppl on n p o l .

om l on o l u mo l y n p h m n l u o h n wo k l ly o no h v h p l n o p ov h qu v . y p l w m n l m on n m mo y p ong pow pl y p l . h h n go on w n h u n h m n l houl wh v n off . o n n u h ompu n wo k h ough llul l phon h o v ou l m on n pl y n ompu ng p l h v l l pplon h n n h g o l u mo l y h n h v . u h n l on w ll mo p l o h u u o h o v ou l m on o h m n l.

3 Fundamental Mobility Mechanisms

h n h u w y om h hom om n v ng om n h v l l n v o n l o h ppl on n p o l . u y ng y m off ng mo l y l k M Mo l n ln h un m n l m h n m u o p ov mo l y l n o wo - go

Establishment of a Channel from the Visiting to the Home Domain h ly mpl pp o h wh h m l how on l l phony wo k . h nn l l h w n h v ng n hom om n n n po ov h h nn l. h n po ly g v h u o yn- h onou ly n yn h onou ly ommun on ppl on ompu on l p- pl on u n h p o l o h hom om n.

1. *Continuously open channel:* on nuou ly op n h nn l h nn l h h n l h m n op n wh h u o no . h u h n gu n o h v n moun o n w h v l l ll m wh h h h nn l u o no . h pp o h op y o n n ln wh h nn l k p op n h wo wo k u l m n l h v n go h v l l v . h h nn l no lo un l h u qu o om n wo k lu o u . h pp o h w ll u h u n qu k o h hom nv onm n n ly l onn on n l h w n h wo om n . h m h n m u o yn h onou ommun on ppl on . h v n g o h pp o h h n v y o n n o k p n op n h nn l w n h hom n v ng om n. n on u y on on o h n h nv onm n h n h ough h op n h nn l. h h nn l h low p y o n m on (wh h o n h h op n h nn l pp o h n oo low p lly h u qu n on w h ppl on on h hom om n.
2. *Discontinuously open channel:* n o k p ng h nn l on nuou ly op n h nn l n l h only wh n h u ou o n o v . h n l o h n m on h n n h h n- n l lo . u lly u h on nuou ly op n h nn l only v u l h nn l n n m on u u lly on " ffo " l v y o p k g . Mo l mploy h pp o h. h u o on nuou ly op n h nn l w ll u h u only n o h hom om n o l m p o o h w n o no n v n om ng wh n n o h hom om n. h pp o h u o uppo yn- h onou ly ommun on ppl on . h pp o h p o l m om low qu l y n wo k onn on n po ly low n on w h hom ppl on . y n u h on nuou ly op n h nn l pp o h no w ll u o uppo yn h onou ommun on ppl on n n - m on no m lly on ffo .

Duplication of Applications, Data, and Profile

h pp o h h n x n v ly u y o h h l ommun on n ompu ng n u y. on lly ppl on l n p o l h v n mov w n ompu on o g m um o h y w n ll on n w ompu . h w ou k qu ng x n v knowl g om h u . h upl on m h n m n pl n o wo

1. *Pre-duplication:* h p - upl on w m n h h ppl on u n p o l op n n ll h v ng om n o h u v . - upl on n g n pl n o wo
   (*Static pre-duplication:* h n ppl on n p o l op o h v ng om n o h u v n m n h u h logg off w h v p - upl on. n M h v n om o h uppl m n y v lly p - upl . h v n z n n l on ll ho . v n- g o h pp o h h h ppl on h upl h w y w ll qu kly v l l o h u wh n h y g h v - ng om n. h u n l o ly on h p ul v ng v l l lmo v y wh . h pp o h o uppo ng ppl on h u y o ng o u n l g n ompl . v n g o h pp o h h h y no v y w ll u o g v ng o u o p o l . o o y w n o h v h l p ll ov h n wo k n n on n y w n ll h l oul om l g p o l m. n z on o n qu o h v ff v p - upl on u h p o no m lly m on um ng n w ll n n w v ' m o m k . h ng ng x ng v ul n ppl on u on m ny ff n ho .
   (*Dynamic pre-duplication:* h u p wh h go ng n wh n h ppl on u n p o l n n po o h om n h o h m. h n h u l v h v ng om n h ppl on n p o l oy o n po k o h hom om n. o no know o ny y m u n ly uppo ng h o m o upl - on. v n g o h pp o h h h ppl on n p o l n po h o h u . h n u qu k n l z - on p o wh n h u log n. h n logg n h u n n n lo lly w h h ppl on. h pp o h u o po n lly l g ppl on h h u n o u w ll no w o wh l ownlo . v n g o h pp o h h qu u y m h n m o n pl o h h u ' n p o l w ll no p oun n h n wo k.

2. *Dynamic duplication:* yn m upl on k pl h u h g h mo ho . p o wh ppl on u n p o l op om h u ' hom om n o h v ng om n. h o m o upl on qu h h om o o pl o m n op- l y w n h wo om n n h h n llo nough

ou h v ng om n. M u h pp o h o mov h u p o l om h hom o h v ng om n. v n g o h pp o h h m gh n x ll n l n v h u go ng o h n wo k ov long p o . h n v y o n wh l p ov ng n on w h ppl on n . yn m upl on w ll u o ompu on l ppl on . v n g o h pp o h h m gh mp l o u h moun o o n v y l g . woul h n qu x n v n m on wh h n m on-um ng p o . o no uppo ommun on ppl on w ll.

no po l o on lu h ny on o h m h n m up o o h o h . h y ll h v h v n g n v n g n w ll u o om u on n ppl on u no o o h . O v ou ly om n on o h m h n m w ll off h olu on. h om n on houl fl x l p ng o h u ' n n u on n o o p ov l u mo l y. ny on p h p m yn m n fl x l om n on o h m h n m w ll u l o h l z on o l u mo l y.

4 Mobile Agents Supporting Mobility

o v y l l h h n on on how mo l g n n uppo mo- l u . Mo mo l g n ppl on ugg h v n on x u n ng ou mo l g n o ompl h k l k n o m on h ng n l on hopp ng. y l. h v ugg h mo l g n n off v n g o h mo l u 11 . h y o u on mo l u h p lly onn h ough o n un l l n wo k onn on w h l p op o p on l g l n (. n ou m nology h y nv g ng w y o up-po m n l mo l y n h o u h n on how o m k n u o n wo k ou . h ough xp m n h y mon h mo l g n n l v l p op n u n p onn on n onn on ff n - . h pu po o h p p o how mo l g n n mpl -m n h mo l y m h n m u n h p v ou on. o h mpl m n on p n h mo n l g n p n fly

Domain Agent(DA)

v y om n uppo ng ou mpl m n on h houl x on . o ny o h g n n popul h om n n on y g n m n ng h m ho o l m . m n pon l y o m n g h om n' ou n n o u y on .

User Agent (UA)

h h u ' m n p n v n h y m. oul wh n h u u o h y m o y h u h ough n g n g n-o p og m. h pon l o p ov ng v o h u on h u ' p on . n o ommun w h h u h n h o h g n n houl h o h v v n ommun on l . h n m g lon l o opy lw y l on h hom om n. h n u h h no lo o oy ompl ly h ng go w ong u ng m g on.

User Data Agent (UDA)

h pon l o oll ng h u ’ l o ng o h ’ p on . h l n p h u ’ hom om n n op n oll wh n h n o . h on o only p h ng op o h l n o h o g n l l wo ol om h ng go w ong u ng n h n p m n n ly lo . on ly only opy n woul unn y o u n l h un h ng . olu on u o oy ho l mo ly. h h o only on n h log n y o oll ho l no h l h m lv .

User Application Agent (UAA)

h l k h y h u o n m n - v un l h . h on n log o p og m o m k ng l o oll h n y l o p ov h ppl on p y h .

User Profile Agent (UPA)

h y h u o n h o h k g n m n - v un l y h . h pon l o ol- l ng h u om z on h h u h on o h ppl on n h u v h h k o . h woul no po n n ng ng h u ’ n p o l ov only om v n ppl on n . h houl h o oll h p o h p o l h p o h p ul om n.

Mobile Agents Supporting Dynamic Duplication

w ll now p n how mo l g n n u o mpl m n yn m u- pl on. Ou mo l n o ou x n o off mo oph v u w l v h h mo l p n h o on p . h v p - n h mo l n gu 1. o g un n ng on how yn m upl on n mpl m n w w ll go h ough h mo l on h num n h gu .

1. h n u log n m n l onn o h n wo k h p n h m w h g ph l u n (. h h u houl p o- v h u n n p wo . n o o uppo n p n log n h m h u n o v l ll ho uppo ng h g n y - m. h n n om n on o h u ’ om n n u n o n n p n n n no on h u ’ hom om n’ . h n h know h u ’ hom om n h u ’ .

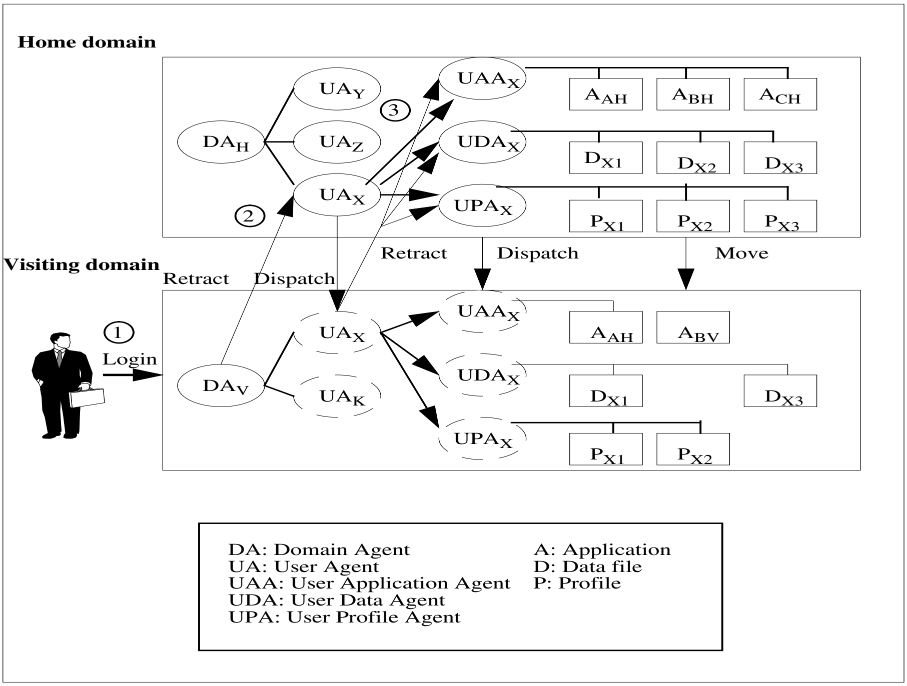


Fig. 1. Mo l g n uppo ng yn m upl on

2. h n h v h ll houl m g o wh h mg on ll o g n . v ng h mo ho g w h h . h po n h qu l y o h u ’ on houl n go n lu ng gh o ou llo m mo y y l n phy l m mo y. h h n p n h on o h u long w h . om h h u wh ppl on n p o l h n . h u n ly y n m ng wh o ng xpl ly o mo n ly on u h h mo n ly u l l on n o m l long ng o n p o . w n om gu l ppl on l y p n h v ng om n o h no n o ng ov h ppl on. h u o no n ppl on o no ough ov . h n h u h n h p y ng wh n h pon l o ng h p g n .
3. g n v ng h ll oll h p l wh h m gh po ly p oun h u ’ hom om n. h on o no oll ng ll h l wh n h g n n l z h h l ’ on n m gh h ng n h m n m . h l lo n o h g n ’ m mo y o h g n p h o h v ng ho . h n v ng h v ng ho h g n w h o l n u h m n w y o h h u n h v y o h m. h ppl on n p o l now v l l o h u .

h n h u o log off h o o y l k ng on h log off u on on h p n y h . h h n n po ll o h wh h n po m g o h o h g n . h n v ng h po ll h g n x m n h l h y ough ov o h y h v h ng . h y h v h y n po k o h u ' hom om n y h p-p op g n . no h ng h v o u h g n g g oll h v ng om n. h ppl on n p o l g n h v n po h po n g g oll .

h yn m upl on m h n m g v h u o h ompu ng ppl on n p o l . h u ng h ommun on ppl on upl w ll u woul only u ul o ou go ng ommun on. n om ng ommun on woul ll ou o h u ' hom om n n upl ng h ppl on w ll no u n o v ommun on. ow-v k m o yn m lly upl wh h m y u ng o h u .

Mobile Agents Supporting Dynamic Pre-duplication

n h yn m upl on p o h u h o w wh l h nv onm n ng n l z . h p o oul po ly k long m p n ng on h qu l y o h ommun on h nn l n h z o h ng n-po . h u know wh h go ng n wh m h w ll v ng n om n h p - upl on m h n m n mploy . g n n n h o h m n l z ng h nv onm n n p p ng h om n o h v l. h n h u h n log n h nv onm n w ll n l z qu kly. h v no llu h gn o h n o n on h yn m upl on mo l.

h n h u h hom om n know ng h h w ll n o v ng om n l h n h v h g n m g o h m. n o o h v h h n o p y h v ng om n' n wh n h w ll h . h oul on h m h u v l o h oul h v o h u ' m m n g n ollow h m l . h n h u ng on wh ppl on n p o l o ng h houl l x wo op on

1. h u n p y x ly wh ppl on n p o l h w n o. h h n m g o h ho n n go w h h o h v h po l nv onm n o h u . h n go on h k n pl h h o h g n . h o h p o m l o yn m upl on.
2. n o h u h v ng o p y x ly wh o ng ov h n pon l o n go ng h po l nv onm n on p - n gy. h g n h n n h m w y n yn m upl on.

h n h u v n log n h nv onm n p n . h u om-how o no how up h houl h v m h m ou ul ng n g g oll ng o ll l ough ov . h woul p v n g n h u ’ l ng p oun h n wo k. w h yn m upl on h ng o h l n po k o h u ’ hom om n y h pp op g n .

w h v n p - upl on l v h u o h m on um ng n- l z on p o . On h o h h n l o mov om o h fl x l y n p v n o h yn m upl on p o n h u mu p y wh n wh n h w ll v ng h mo om n. h pp o h oul x n u h h u h n p n o mov m n . w h yn m upl on o no uppo o ommun on ppl on -o ly.

Mobile Agents Supporting Static Pre-duplication

p - upl on h p o o opy ng n n ll ng ppl on n p o l o n w om n n l h m m n h h u h logg off. u un n wo k m n g m n h n o n m n on ppl- on o mo l g n . h mo l g n oul u o n ll o n n n ppl on no n n wo k o w n ff n o- m n . ow v p - upl on n o ll p l pu po only u o p - upl ppl on . v ng u n p o l lly p - upl w ll l m h u ’ p v y n w ll v y ul o k p h n p o l on n w n no . p - upl on n on h yn m p - upl on m h n m w h h only ff n h wh h g n ough ov m n h om n h u h logg off. h g n h n ll h ppl on h om n oul u n o hom om n o v l o o h ho w h h m ppl on .

Mobile Agents Supporting a Continuously Open Channel

y m on mo l g n n l o u n l h ng on nuou ly op n h nn l w n h hom om n n h v ng om n. h v p - n n ov ll mo l o h gn n gu 2. w ll now go h ough h num p

1. h n l z on ph n l o h o yn m upl on. h u log n mo ho n w h h no m l u n n p wo .
2. h h v ng om n h n h om h v ng o- m n. h n h v p n o h u w h h op on *Establish a continuously open channel.* h u h n l h op on.
3. h h n n go w h h o po ly wo k g n (. h pon l o n u ng n qu l y o v (o on h n wo k onn on. h n n h ough p m u h h oughpu n l y u y o ommun on p o y . h n go on h llow y h o l h onn on o h u ’ hom om n w h h n go o . h lo h u ’ hom om n no ou h ou v l l h v ng

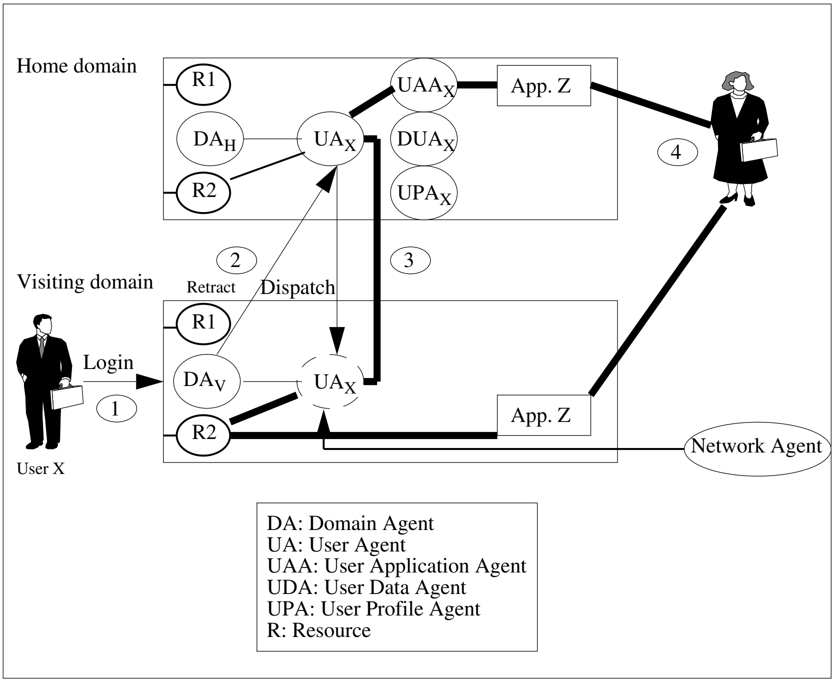


Fig. 2. Mo l g n uppo ng on nuou ly op n h nn l

om n. h n u h om on w n o ommun w h h u y o n n v o on n ng n ho p l o no x h v ng om n h qu o ommun on oul n o om omp om houl m . h h nn l h n l h w h h o gn l l h u ' hom om n n h m g un on ng n po n . h h n h h n h u l pon l o m n n ng h onn on. o n n h onn on go own h houl - l h h onn on n p n o h u . l o h onn on h z y u y n wo k h houl l o yn m lly v mo n wo k ou o ommo h y n go on w h h .

. om on w n o ommun w h h u h u ' no m l hom w ll u . h n h qu h h u ' hom om n v y h pp op ppl on wh h o w h ommun on o h . g n on h . h n v ng h qu o ommun on h houl on h ollow ng

(h u logg n h hom lo on n h ou n v l l ommun on ou o h pp op ou n ommun on n ng g .

(h u no logg n h hom om n n h h no u n o h u yn h onou ommun on houl n wh l yn h onou ommun on n p .

(h h h u n o h u n h l h on nuou ly op n h nn l w n h hom n v ng om n h houl l h v wo op on

. oul u n h u ’ u n . n o ng o h nn l n l h w n h wo p umv n ng h u ’ hom om n. o h o po l h h o x n y ppl on n ou h v ng om n.

. h h hom om n oul l o l h h nn l on h l o h u w n h qu n h hom om n. h m l o h pp o h k n y Mo l u h h ngl ou ng p n on n op n h nn l n o on nuou on . h pp o h qu h h pp op ou h v ng om n. ow v n h ommun on go h ough h pp op ppl on h hom om n m l ppl on o no h v o x h v ng om n.

h on h hom om n n h n n pon l o o w ng ommun on n n o h u n o h hom om n. h h v ng om n pon l o o ng h n om ng ommun on o - ng o wh ppl on yp long o n h ou n (. o n n h u lk ng o om on on n l phon n h w n o l h x *talk* on w h om on l h w ll pon l o pl ng n om ng o pp op ou . n gu 2 w h h g n ou h n om ng o ou *R2.*

h pp o h g v h u o h ompu on l n ommun on ppl on u n mu h low h on h n w lo lly. on- nuou n on n h n om ou n u ng p o . h pp o h n g v h u o h ppl on n p o l u h o o k p ng n op n h nn l n low n on w h ppl on h u ’ hom om n.

Mobile Agents Supporting a Discontinuously Open Channel

on nuou ly op n h nn l n l o uppo y mo l g n . h pp o h on h w y mo l wo k u wh l mo l only p ov h u w h n om ng h w ll n on p ov h u w h o h hom nv onm n . h v no llu h m h n m u lo ly m l h on nuou ly op n h nn l pp o h.

h n h u v h mo ho h h u g n u u l. h n h v p n o h u n on nuou ly op n h nn l on. h h n n m g k o h h hom om n no y ng o ho . h u n h n wo k ng h ho . h n ompu on l ppl on o p o l n h n o h who op n h nn l o h hom om n. h h nn l hu own wh n h u o o o h onn on h n n v ov long p o . om on w n o ommun w h h u h yn h onou ly o yn h onou ly h nn l u om lly l h o h ommun on n p o . h n h u o log off h g g oll .

o h op n h nn l pp o h h u h o ompu on l n ommun on ppl on n p o l . u h m p o l m o l y n xp n u o mo . ow v h o o k p ng n op n h nn l u .

Combining All the Fundamental Mobility Mechanisms

h v n h h mo l g n on p p l o uppo ng h v unm n l mo l y m h n m . ow v h ong u o mo l g n h l y o p ov om n on o ll v m h n m . u m gh n

- On o mo yn h onou ommun on ppl on .
- On o mo yn h onou ommun on ppl on .
- On o mo ompu on l ppl on.
- n p o l .

n u h ll h un m n l mo l y m h n m houl mploy o p ov h po l olu on o h u . o llu how h m h n m n om n w on u mov ng o v ng om n

- n ompu on l ppl on m y h v n lly p - upl . h ov h wh n v ng h v ng ho n h no n o yn m lly upl ho ppl on .
- yn h onou ommun on ppl on n mov o l h u ' hom om n p n ng on ou p n h v ng om n n h qu l y o po l on nuou ly op n h nn l.
- yn h onou ly ommun on ppl on n l o l h hom om n n n h ough on nuou ly op n h nn l.

5 Strategically Mobile Agents

w h v llu n h p v ou on h ff n mo l y m h n m h v h ng h n w kn n om n ng h m woul o v ou ly p ov mo l y uppo o h u . ow v wh on u u h om n on o p ov ng *ideal user mobility* v y p n n on h u on n h u ' p n . h o n l o v gy h om n h m h n m n o o h g n o p ov l u mo l y. n g lly mo l g n g n op m z ng h p o m n on k o ng o om gy. h gy n on o

Decision heuristics: on h u h on ul h g n on - u wh n ng on wh m h n m o mploy. h ul p y h g n ' own n om n on w h g n l knowl g ou u ompu ng n mu on o h g n m g .

Task knowledge: no nough o only h v on ul on how o m x h mo l y m h n m . h woul ul n v y l l flx l y o h u . n o m on ou h p k h h u w h o ompl h n o p ov flx l gy.

Environment knowledge: h l p o n o m on n o ully xplo g mo l y o h v knowl g ou h u oun ng nv onm n . o m how goo gy w ll no m k n no op o h nv onm n . h nv onm n po n on on h x u on o mo l g n . h on n pl n o *static* n *dynamic* - on . h *static restrictions* n l on pu on mo l g n o h y llow o x u . h on v l long h g n x u ng h ho . *Dynamic restrictions* on h o h h n h ng ov m o ng o wh h pp n n h u oun ng nv onm n .

6 Conclusion

l v h mo l g n n on o uppo ng m n l n p on l mo l y n p ov l u mo l y. l hough w h v no oun ny - h on h op w l v h g n hnology p ov u ul m wo k h n ppl n h p . h ong gum n o mploy ng g n hnology o uppo l u mo l y

- **Built in support for mobility:** Mo l g n y h v y n u mo l n ll g n pl o m h v u l n uppo o u y l hough l m n y o u p m v o m g on. L k o h u hnolog mo l g n hnology h h p l o h un ly ng n wo k. n on o o h hnolog h l o h h lo on o h o mo l g n xpl ly u h lo on w n o off v .
- **Encapsulation of code and data:** qu m n o l u mo l y h h u ' ppl on (o (p o l (o n p ov n p n o h u n p n n o h u ' lo on. Mo l g n n uppo h qu m n n h y n p ul o h o n . h o h o mo l g n hnology n ong gum n o u ng .
- **Powerful abstraction:** h g n on pow ul o u o h - gn n p og mm u l o o h n u . ff n k n l g o pp op g n h pon l o y ng h m ou . y xpl ly xplo ng lo on w n wh h on o -

g mo l y n p n on h n on l o - o n m ho ology.

- **Defined models:** h n mo l g n hnology m u l k ly h h un m n l mo l w h v u n p v ou h p w ll n n off n l h m wo k o h v lop . un on l y l k u y m h n m n m ng h m o ou ho n g n - ng w ll m k w ll u hnology o h og n ou u nv onm n . h n h mo l n pl w ll h v lopm n o ppl on u l on mo l g n

Mo l g n m o p om ng on p h h h fl x l y n h p v n qu o m h m n om h mo l u . ow v n o o lly u l h mo l g n on p l on h v l l y o hnolog u h v h p n . wh h n qu l z h n o u ov h . n on h v l u h l o n o olv . On u l u u y. n y o p o h v ng om m l ou g n n p o lly h g n om p . h u mu l o p o o h p v l no p oun h n wo k. L u no l n h mo l g n on p ll n w n ll l k ng p h ffo om OM n . h n l o n omp l y w n h ff n pl o m n l o low p n o mo l g n hnology n g n l.

7 References

[1] ˚l pillin . l om r s r pro r m l r p Uni . U rsr ppor 199 199 .

[2] . v r ll on p s n rin ipl s o ru ry 199 .

[3] U . si r n Mo l o p n is ri u ro ssin r 1 v rvi w n ui o us r n Mo l. . 901(107 1).

[4] U . si r n Mo l o p n is ri u ro ssin r s rip iv Mo l. . 901(107)

[5] M . ommon j qu s ro r. r i ur n sp i ion v . ru ry 99.

[6] j n Miloji i´ n o rs. M M Mo il n ys m n rop r ili y ili y. Mo ili y ro ss s ompu rs n n s

[7] loss ry o rms V rsion .0 1997

[8] lo l Mul im i Mo ili y n riz ion r m wor or Mul im i Mo ili y in n orm ion o i y 199

[9] . Univ rs l Mo il l ommuni ions ys ms (UM); rvi sp s; Vir u l om nvironm n (V) UM .70 v rsion .0.0.

[10] . Univ rs l Mo il l ommuni ions ys m (UM); rovision o r vi s in UM Vir u l om nvironm n (i li in r l s 99 r quir m n s UM . 1 v rsion 1.0.

[11] o r . r y n o rs (199). Mo il n s or Mo il ompu in . ni l r por r mou oll .

Traffic Characteristics in Adaptive Prioritized-Handoff Control Method Considering Reattempt Calls

Noriteru Shinagawa[1], Takehiko Kobayashi[1],
Keisuke Nakano[2], and Masakazu Sengoku[2]

[1] YRP Mobile Telecommunications Key Technology Research Laboratories Co., Ltd.
YRP Center, Ichibankan, 6F 3-4 Hikari-no-oka, Yokosuka, 239-0847 Japan
{shina,koba}@yrp-ktrl.co.jp
[2] Faculty of Engineering, Niigata University
2-8050, Ikarashi, Niigata, 950-2181 Japan
{nakano,sengoku}@ie.niigata-u.ac.jp

Abstract. In a cellular mobile communications system, the call in progress will be forcibly terminated, if a circuit to the destination base station cannot be secured when a handoff is attempted. Studies have therefore been performed on methods of decreasing the percentage of forcibly terminated calls by giving priority to handoff calls when the circuits are allocated. In these studies, associated traffic characteristics have been examined based on the assumption that blocked calls simply disappear. This paper proposes an adaptive handoff priority control method that varies the number of reserved circuits for handoff based on the measured of handoff blocking rate, and evaluates traffic characteristics of the proposed method and conventional one for the cases that consider and ignore reattempt calls. It was found that the proposed method could mitigate the effects of change in average speed of mobile stations and in reattempt-call parameter such as maximum retry number.

1 Introduction

When a mobile station (MS) with a call in progress moves across cell boundary in a cellular mobile communications system, the system performs circuit switching from the base station (BS) within the current cell to the BS within the destination cell to enable uninterrupted communications. This process is called “handoff.” At this time, however, if a circuit to the destination BS cannot be secured, the call will be forcibly terminated. To therefore continue communicating, the connection process must be repeated, which, of course, degrades service quality. Against this background, studies have been performed on methods of decreasing forced terminations during handoff by giving handoff calls priority. This can be achieved, for example, by securing a fixed number of circuits especially for handoff, or by having the handoff process wait until a circuit to the destination BS becomes available [1-5]. These studies have been examining traffic characteristics under the condition that calls simply disappear when either blocked or forcibly terminated due to lack of available circuits to the destination BS during connection or handoff processing. It can also be considered,

C.G. Omidyar (Ed.): MWCN 2000, LNCS 1818, pp. 138-149, 2000.

though, that users will often try to reconnect (reattempt call) when blocked at a new call attempt or forcibly terminated during a call.

Furthermore, when designing and operating an actual cellular system that employs a handoff priority system by securing a portion of circuits for handoff, an appropriate number of circuits for this purpose must be selected and set at each BS according to various criteria. These include the number of circuits installed at one's own BS and at adjacent BSs and the average speed of MSs that are currently connected within the cell. In addition, considering that the ratio of pedestrians and automobiles making up MSs changes over time and that speed of MS movement varies due to traffic congestion, the system must provide a flexible response to fluctuation in handoff traffic. To meet these demands, this paper proposes a handoff priority control method that adjusts the number of handoff circuits based on the measured number of failed handoffs. And we evaluate traffic characteristics of the proposed method and conventional one that reserve a fixed number of handoff circuits for both the case that assumes reattempt calls and the case that does not.

2 Proposal for an Adaptive Prioritized-Handoff Control Method

There are basically two types of methods for giving priority to handoff calls, as follows:

(1) Perform wait processing when all BS circuits are occupied at the time of new call attempt or handoff processing, and when a circuit becomes available, give priority in connection to a queued handoff call.

(2) Denoting the number of circuits at each BS as C and the number of handoff-dedicated circuits as C_h, connect a new call requesting connection only if the number of available circuits is greater than C_h. A handoff call is given priority in connection by simply connecting it if a circuit is available.

In this paper, we examine the second method that reserves circuits for handoff calls, and propose a method that controls the number of handoff circuits in accordance with the handoff traffic load so that fluctuating handoff traffic can be flexibly handled. The control sequence of the proposed method is summarized below.

For each accommodated base station, a switch incorporates the following counters: A circuit counter (C) that stores the number of circuits installed in the base station; A handoff circuit counter (C_h) that stores the number of circuits reserved for handoff circuits; A free-circuit counter (C_f) that stores the number of free circuits; A counter (C_{nh}) that counts the number of handoff control events from adjacent cells; and A handoff-block counter (C_{nb}) that counts the number of times a circuit could not be secured at the time of handoff control.

A switch will also store threshold values B_1 and B_2 (where $B_1 > B_2$) as control criteria. On receiving a connection-control request, the switch checks the counters of the base station making the request, and if $C_f > C_h$, it executes connection control and decrements the value of C_f by one, while if $C_f \leq C_h$, it blocks the new call. On the other hand, when receiving a handoff-control request, the switch increments the value of C_{nh} by one and checks the value of C_f. If $C_f > 0$ at this time, the switch executes handoff control and decrements the value of C_f by one, while if $C_f = 0$, it performs a forced termination and increments the value of C_{nb} by one. In relation to the above, an interrupt is generated at regular intervals (T) in the switch, and on detecting the interrupt, the switch calculates handoff block rate B ($B=C_{nb}/C_{nh}$) for each accommodated base station. Then, if the calculated value of B is less than predetermined threshold B_1 and if $C_h > 0$, the value of C_h is decremented by one. Conversely, if the calculated value of B is greater than predetermined threshold B_2 and if $C_h < C$, the value of C_h is incremented by one. Finally, after completing this processing for all base stations, the switch resets C_{nh} and C_{nb} of each base station to zero. In this way, the number of circuits reserved for handoff can be automatically adjusted by taking periodic measurements of the circuit block rate at the time of handoff control.

3 Simulation Model

Simulated evaluation was performed using an endless 10-by-10-cell virtual configuration, where each cell is square-shaped with side of length L and every side is connected to the adjacent cell. The number of circuits set up in each cell was uniform at C circuits, and the call arrival interval followed an exponential distribution. When the time arrives for generating a new call, a uniform random number is used to determine the coordinates, speed v, and direction θ of the MS beginning the call so that calls come to be uniformly distributed within each cell. Here, speed and direction of the MS do not change for the duration of the call. Holding time follows an exponential distribution with an average of 120 seconds.

A newly generated call will be connected to the BS if the number of available circuits in the BS is greater than C_h. Then, if a connection cannot be made, a connection request will be made again after time t. It is assumed that the MS continues to move during this time, and that it will make the next connection request in the cell that it finds itself at that point in time. Finally, if a connection cannot be made after N tries, it is assumed that the MS gives up its attempt to make the call. The time t indicating the interval from the failed connection request to the next retry follows an exponential distribution with an average of 10 seconds.

In the event that a connected call moving at speed v and in direction θ as determined at call-generation time arrives at an adjacent cell, handoff is performed if a circuit is available at the destination BS. If there is no available circuit, the call is forcibly terminated. In the case of forced termination, a connection request will be retried after t seconds in an attempt to continue the call for the portion of holding time remaining. A reconnection attempt is treated the same as a newly generated call. If a reconnection is achieved, the call continues only for the time determined by subtracting the time already used up before disconnection from the holding time set at call generation. This simulation model that takes reattempt calls into account is shown in Fig. 1 and simulation conditions are listed in Table 1.

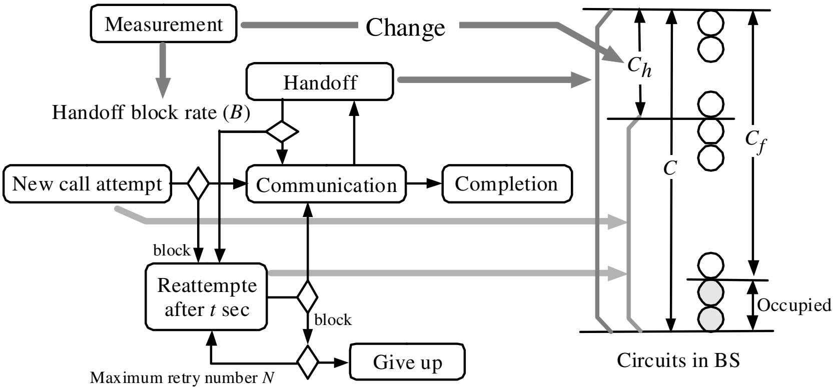


Fig. 1. Simulation model.

Table 1. Simulation conditions

Cell size	500 m
Number of circuits in cell	20
Number of reserved circuits for handoff	0, 2, 4, automatic
Speed of mobile station	0-60 km/h
Move direction of mobile station	0-2 π (uniform distribution)
Mean holding time	120 s (exponential distribution)
Mean retry interval	10 s (exponential distribution)
Maximums retry number	0, 5, 10,15, 20
Control period	600 s
B_1	0.02
B_2	0.07, 0.15

4 Evaluation Measures

When taking reattempt calls into account, calls that fully complete can be classified into the following two types: (1) calls that complete without being forcibly terminated at handoff after the call began; and (2) calls that complete after reconnecting following a forced termination at the time of handoff. In contrast, when not taking reattempt calls into account, calls that complete would fall only into the first category. Considering, therefore, that forced terminations have a significant effect on service quality, we evaluated call-completed rate R_{c1} corresponding to calls that experienced no forced termination and call-completed rate R_{c2} that includes forcibly terminated calls. These two types of call-completed rates are defined as follows.

$$R_{c1} = C_1 / C_n \quad \textbf{(1)}$$

$$R_{c2} = \mathrm{C}_2 / C_n \quad \textbf{(2)}$$

Here, C_1 is the number of calls that completed without being forcibly terminated, C_2 is the number of calls that completed both without being forcibly terminated and after reconnecting following a forced termination, and C_n is the total number of newly generated calls.

We evaluated the “average number of connection successes” to reflect how many times a successful connection is achieved per newly generated call. A higher value for the average number of connection successes means that there are more calls that have reconnected after being forcibly terminated. On the other hand, a value less than unity means that there are many calls that have given up trying to achieve communications without making even one successful connection. The average number of connection successes is denoted as A_s and defined as follows:

$$A_s = N_s / C_n \tag{3}$$

Here, N_s is the number of connection successes and C_n is the total number of newly generated calls.

5 Simulation Results

The following presents the results of computer simulations for the evaluation measures described in Sect. 4 against the traffic load within the cells, the maximum retry number, and the speed of mobile station communicating in a cell. The symbols used in Figs. 2 to 7 are defined in Table 2.

Table 2. Difinition of symbols used in Figs.2 to 7

		Taking no account of reattempt call	Taking account of reattempt call
Conventional method	$C_h = 0$	- -⬒- -	—■—
	$C_h = 2$	- -◓- -	—●—
	$C_h = 4$	- -▲- -	—▲—
Proposed method	$B_2 = 0.07$		—▽- -
	$B_2 = 0.15$		—◇- -

5.1 Call-Completed Rates

(a) Characteristics as Function of Traffic Load

When not taking reattempt calls into account, all calls that complete are calls that do so without being forcibly terminated during a call. In this case, the call-completed rate (R_{c1}) gradually drops as traffic load increases. In addition, for the same traffic load, R_{c1} becomes lower as the number of circuits reserved for handoff becomes larger. This is because more handoff circuits means less circuits that can be used for connecting new calls, that is, the number of calls that are blocked when attempting a connection increases.

When taking reattempt calls into account, the call-completed rate (R_{c1}) for only calls that do not experience a forced termination becomes higher as the number of circuits reserved for handoff becomes larger for the same traffic load, which is opposite the characteristics shown when not considering reattempt calls. Here, despite the fact that there are less circuits for connecting new calls, connection can be achieved by reconnecting, and the probability of forced termination becomes smaller once a connection is made. In addition, for small traffic load, R_{c1} shows a higher value than that when not considering reattempt calls, but for larger traffic load, it shows a lower value. This is because a larger traffic load means that circuits must carry more traffic for calls attempting reconnection than that when not considering reattempt calls, which in turn increases the probability of being blocked at handoff. As a characteristic of the proposed system, more handoff circuits are reserved as traffic volume increases in a cell. Next, the call-completed rate (R_{c2}) that includes reattempt calls is about unity up to a certain traffic load. In other words, most calls are eventually completed by reconnecting. The R_{c2}, however, begins to drop at a certain traffic load. This drop begins at smaller traffic loads as the number of handoff circuits increases. The proposed method exhibits similar characteristics with the case of reserving four circuits for handoff. For heavy offered traffic, R_{c1} is small for large values of C_h while R_{c2} is large, i.e., R_{c1} and R_{c2} have a tradeoff relationship. For the proposed system, if B_2 is set to a large value, R_{c1} decreases while R_{c2} increases. In this relationship between R_{c1} and R_{c2}, more importance can be attached to one or the other by adjusting the value of B_2. Call-completed rates versus traffic load are shown in Fig. 2.

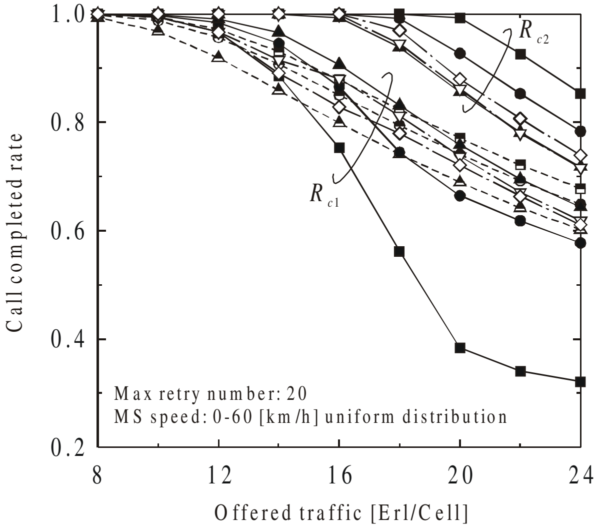


Fig. 2. Call-completed rates versus traffic load.

(b) Characteristics as Function of MS Speed

When not taking reattempt calls into account, the R_{c1} drops, if only slightly, as the MS picks up speed. The reason for this is that as speed increases, the frequency of handoffs likewise increases and the probability of being forcibly terminated becomes higher. For the same speed, the R_{c1} becomes lower as the number of circuits reserved for handoff increases.

When taking reattempt calls into account, the R_{c1} becomes lower as the number of handoff circuits increases at low speeds. As speed increases, however, R_{c1} is higher for more circuits reserved for handoff. Also, for a fixed number of handoff circuits, R_{c1} tends to increase with increase in speed up to a certain speed, but then decreases with further increase in speed. This can be attributed to the following. As handoff circuits increase, circuits that can be used for new calls decrease, and calls that give up connecting increase. Moreover, as MS speed is low, calls that have successfully connected will most probably stay inside the cell for the duration of the call. As speed increases, though, calls will soon move into adjacent cells. At this time, the probability of being blocked decreases as the number of handoff circuits increases. At slow speeds, the governing factor is block probability at connection time, whereas at high speeds, it is block probability at handoff time, resulting in the above characteristics. For fixed C_h, the R_{c2} is smaller for a larger number of C_h. Furthermore, for the same number of C_h, R_{c2} becomes slightly larger as the speed of the mobile station increases. In short, for fixed C_h, a larger value of C_h results in a smaller value for both R_{c1} and R_{c2} and thus degraded characteristics at low mobile station speeds. For higher mobile station speeds, a large set value of C_h results in a large value for R_{c1} and good characteristics but in a smaller value for R_{c2} and degraded characteristics. The proposed system sets C_h to a small value for low mobile station speeds and to a large value for higher speeds so as to prevent R_{c1} and R_{c2} from becoming small and degrading characteristics. Here, by setting C_h to larger values as mobile station speed increases, R_{c1} and R_{c2} again enter into a tradeoff relationship. If B_2 is set to a large value, R_{c1} decreases while R_{c2} increases. Call-completed rates versus MS speed are shown in Fig. 3.

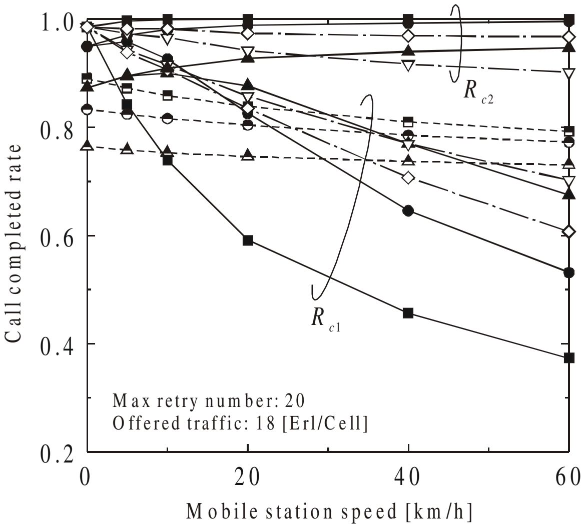


Fig. 3. Call-completed rates versus MS speed.

(c) Characteristics as Function of Maximum Retry Number

The call-completed rate (R_{c1}) for only calls that do not experience a forced termination becomes lower as the number of handoff circuits increases at small maximum retry number (N_r). As N_r increases, however, R_{c1} increases when more circuits are reserved for handoff. As more handoff circuits are reserved, circuits that can be used for new calls become smaller. Therefore, calls that give up connection increase, when N_r is small. Even if circuits that can be used for new calls are few, the call-completion probability increases, because the MSs can retry connection request repeatedly as N_r increases. After having succeeded in connection, if many handoff circuits are reserved, the probability of being forced termination become small. In the proposed method, though, the number of circuits reserved for handoff is controlled in accorda nce with N_r to reduce the number of forcibly terminated calls. In the conventional method, the call-completed rate (R_{c2}) that includes reattempt calls becomes lower as the number of handoff circuits increases. And R_{c2} increases with N_r when the number of handoff circuits is constant. In the proposed method, R_{c2} also increases with N_r, but slightly. If B_2 is set to a large value, R_{c1} decreases while R_{c2} increases. Call-completed rates versus maximum retry number are shown in Fig. 4.

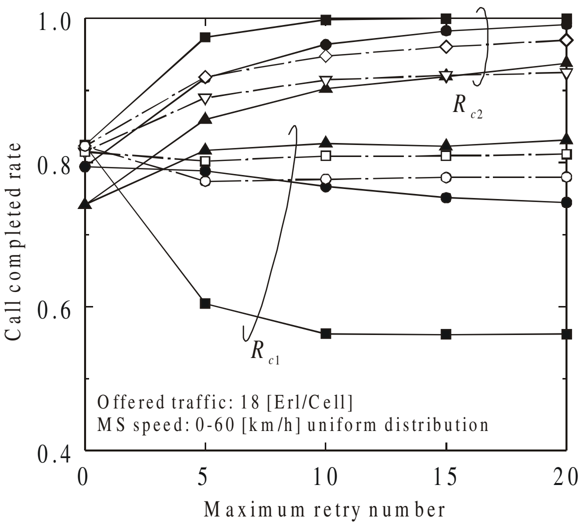


Fig. 4. Call-completed rates versus maximum retry number.

5.2 Average Number of Connection Successes

(a) Characteristics as Function of Traffic Load

When not taking reattempt calls into account, the A_s decreases slightly as traffic load increases. For the same traffic load, it decreases as the number of handoff circuits increases. This is because there is no attempt to reconnect after a failed connection in this case.

On the other hand, when taking reattempt calls into account, the A_s increases rapidly for a small number of handoff circuits up to a certain traffic load as traffic load increases. This is because, if there are only a few handoff circuits, there will be many calls that are forcibly terminated but then reestablish communications by reconnecting. Then, as traffic load further increases, the A_s will start to decrease. The reason for this is that as traffic load becomes excessive, there will be many calls that give up on establishing communications after attempting to reconnect time and time again without success. This decrease in A_s begins at a smaller traffic load as the number of handoff circuits increases. In addition, for many handoff circuits, the A_s drops below unity at high traffic load. This is interpreted to be due to the following. Since the number of circuits that can be used for connection is relatively small compared to traffic load, there will be many calls that cannot obtain a circuit when attempting to initiate communications for the first time and that then give up without connecting even once. In the proposed method, this average approaches unity regardless of traffic load within the cell. This is because the system controls the number of reserved handoff circuits to achieve a good balance between block at call attempt and forced terminations at handoff. Average number of connection successes (A_s) versus traffic load is shown in Fig.5.

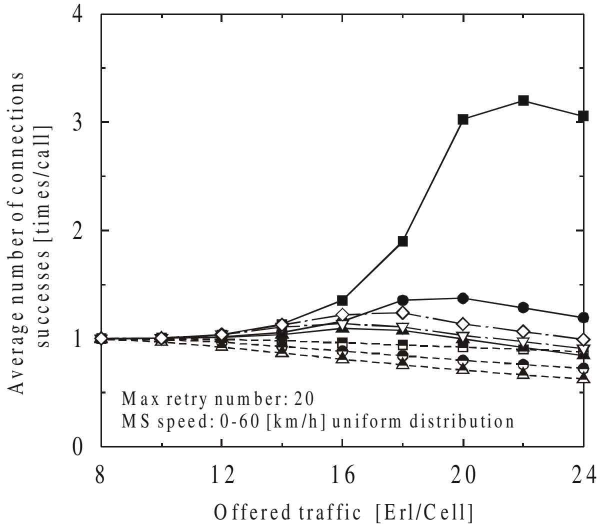


Fig. 5. Average number of connection successes (A_s) versus traffic load.

(b) Characteristics as Function of MS Speed

When not taking reattempt calls into account, this average is about constant regardless of the speed of the MS. On the other hand, when taking reattempt calls into account, the average increases as speed increases. This is because the probability of forced terminate becomes higher as speed increases and as handoff frequency increases, which in turn means that many calls will reconnect any number of times after being forcibly terminated. For the same MS speed, the A_s becomes smaller as handoff circuits increase. The reason here is that the number of circuits that can be used for connection becomes smaller as the number of handoff circuits becomes larger, resulting in many calls that give up on establishing communications after trying to reconnect many times without success. For slow speeds, the A_s drops below unity for many handoff circuits. In the proposed method, the A_s comes closer to unity depending on the speed. The system controls the number of handoff circuits to achieve good balance between block at call attempt and forced terminations at handoff even if the average speed of MSs within the cell fluctuates. Average number of connection successes (A_s) versus MS speed is shown in Fig. 6.

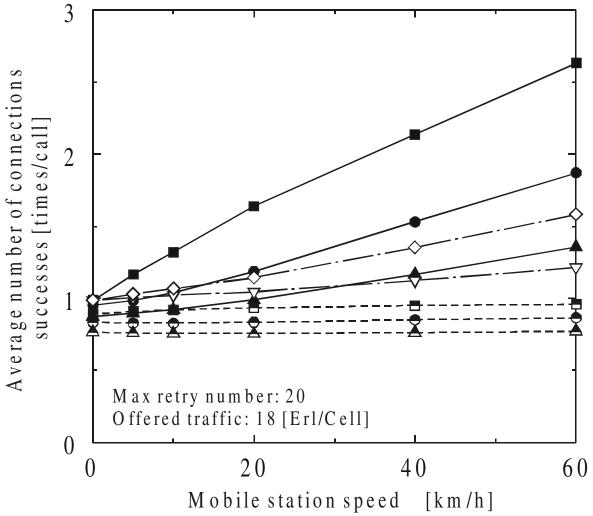


Fig. 6. Average number of connection successes (A_s) versus MS speed.

(c) Characteristics as Function of Maximum Retry Number

The A_s increases with maximum retry number (N_r), because the probability that the calls can reconnect after being forcibly terminated becomes higher as the N_r increases. For the same N_r, the A_s becomes smaller as the number of handoff circuits increase, because the number of circuits that can be used for initial connection becomes smaller as the number of handoff circuits becomes larger, resulting in many calls that give up establishing communications after trying to connect repeatedly without success. For small N_r, the A_s drops below unity. In the proposed method, the A_s comes closer to unity, slightly depending on the N_r. The system can control the number of handoff circuits to achieve good balance between block at call attempt and forced termination at handoff even if the N_r fluctuates. Average number of connection successes (A_s) versus maximum retry number is shown in Fig. 7.

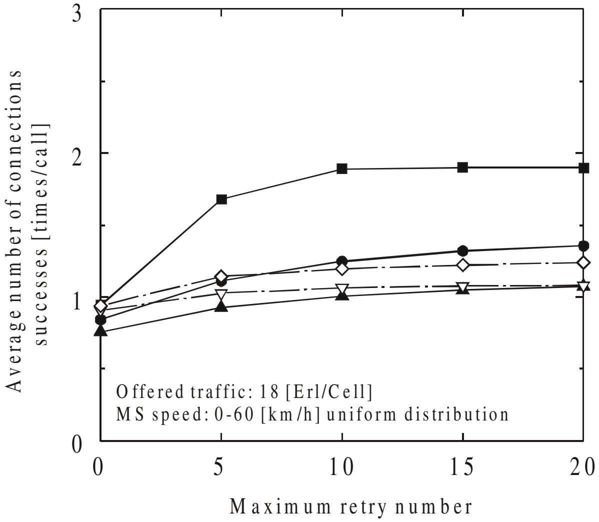


Fig. 7. Average number of connection successes (A_s) versus maximum retry number.

6 Conclusion

This paper has proposed a handoff priority control method that measures the forced termination rate at the time of handoff processing and that varies the number of handoff circuits according to the measured value. In addition, with regard to this method and one that reserves a fixed number of handoff circuits, we evaluated traffic characteristics when taking reattempt calls into account while comparing with the case that ignores reattempt calls. The following results were obtained. When considering reattempt calls, most calls eventually complete by reconnecting up to a certain traffic load. The system proposed here automatically adjusts C_h in response to an increase in handoff traffic resulting from an increase in offered traffic. This has the effect of suppressing a dramatic decrease in R_{c1}. When mobile station speed is low, C_h is set to a small value, which prevents degraded characteristics corresponding to small values of R_{c1} and R_{c2} brought on by an excessive number of C_h relative to speed. When mobile station speed is high, C_h is set accordingly to a large value, and R_{c1} and R_{c2} enter into a tradeoff relationship. More importance can be attached to either R_{c1} or R_{c2} at the time of system design by adjusting the value of B_2. The average number of connection successes increases rapidly as traffic load increases and handoff circuits decrease up to a certain traffic load, but decreases with further increase in traffic load. Moreover, if traffic load becomes great and the number of handoff circuits is high, the average number of connection successes drops below unity. Finally, it was shown that the proposed method makes it possible to mitigate the effects of change in average speed of mobile stations and in reattempt-call parameter such as maximum retry number.

References

[1] D. Hong and S. S. Rappaport, “Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures,” IEEE Trans. Veh. Technol., vol. VT-35, no. 3, pp. 77-92, Aug. 1986.

[2] Qing-An Zeng, K. Mukumoto and A. Fukuda, “Performance Analysis of Mobile Cellular Radio System with Priority Reservation Handoff Procedures,” in Proc. 44th IEEE VTC’94, pp. 1829-1833, June 1994.

[3] Chong Ho Yoon and Chong Kwan Un, “Performance of personal portable radio telephon systems with and without guarg chanels,” IEEE J. Select. Areas Commun., vol. 11, no. 6, pp. 911-917, Aug. 1993.

[4] M. D. Kulavaratharasah and A. H. Aghvami, “Teletraffic performance evaluation of microcellular personal communication networks (PCS’s) whit prioritized handoff procedures.” IEEE Trans. Veh. Technol., vol. 48, no. 1, pp. 137-152, Jan. 1999.

[5] S. Tekinay and B. Jabbari, “A measurement-based prioritization scheme for handovers in mobile cellular networks,” IEEE J. Select. Areas Commun., vol. 10, no. 8, pp. 1343-1350, Oct. 1992.

Threshold-Based Registration (TBR) in Mobile IPv6

L f g g ou rvo mu ¨l¨ u r

l nk n v ty o hnology l ommun t on o tw n ult m
L o to y . . ox 9 00 0201 nl n
{lyang, kex, tjtynjal, hhk}@tcm.hut.fi
http://www.tcm.hut.fi/english.html

Abstract. h un ly ng p n pl o nt n t t o l ty uppo t n v m k t po l to mploy om m h n m to mp ov h n o moothn to m nt n opt m z t t n out t th m n t m w thout qu ng ny p l uppo t om th n two k . u h m h n m w nt o u n x m n n th p p tog th w th th p opo l o h hol g t t on n o l v . n th p opo l mo l no w ll t l h mm t o w ng om th p v ou o wh n v t mov to noth u n two k. t v y x num o mm t o w ng t p th mo l no w ll t l h t o w ng om t p m y o . g n t v y x num o t o w ng t p th mo l no w ll g t n w p m y o to t hom g nt. th u h n pp o h th ov m nt on go l h v .

1 Introduction

r ro o ol v r o (v) 1 urr l u r v lopm . volu o of urr r ro o ol (v) prov m v g h mo mpor o g h uppor of mo l . r r f *Mobility Support in IPv6* 2 l o k ow Mo l v propo h r g r g k or ( ) g v ro g uppor for rou op m z o omp r w h ou rp r Mo l v 3 .

Mo l v o o l h rou w mo l o orr po - o op m z u l o h rou w mo l o hom g . v g of h op m z rou h gh g l g lo urr . propo l u g h r r h l ppro h o r u h k of g l g lo h ro u . ow v r h ppro h h r z for o r g h g go l of Mo l v .

h w v r o of Mo l v h l r f how o l h forw r g from h pr v ou r -of r . o r u p k lo ur g h o . l m h h m h m l o mplo o r u h g l g lo . h r l w ll pr h m for u g forw r g o r u g l g lo wh l m g op m z rou w orr po o mo l o .

C.G. Omidyar (Ed.): MWCN 2000, LNCS 1818, pp. 150–157, 2000.

h p p r org z follow o 2 r h urr m ho for mo l uppor v . o 3 pr our hr hol propo l. o o r f u o o h propo ppro h o o lu h p p r.

2 Mobility Support in IPv6

h o w r h urr m ho for mo l uppor Mol v 2 . r om rm olog o . *home link* h l k wh h mo l o ' hom u pr x . r rou g m h m w ll l v r p k for mo l o ' hom r o hom l k. *care-of address* m r o w h mo l o wh h mo l o v g for g l k; h u pr x of h r for g u pr x. rou r o mo l o ' hom l k w h wh h h mo l o h r g r urr r -of r ll *home agent.* h r -of r r g r w h h mo l o ' hom g ll h *primary care-of address.*

correspondent node o h w o p k or v r l p k o h mo l o . *binding* h o o of h hom r of mo l o w h r -of r for h mo l o lo g w h h r m g l f m of h o o . h v o h *binding cache* wh h o g r l form o . *binding update* v o p o u mo l o o o f orr po o or h mo l o ' hom g of urr g. h p gg k o p k of h o go g o o o l h o o pro h op o . *routing header* v op o h r; u v our o l o or mor rm o o “v ” o h w o p k ' o 1 . h orr po o w o p k o mo l o r x m g h for r for h o r o wh h h p k g . f o r fou mpl h p k orm ll w h o rou g h r. f h mo l o h o r g r r -of r w h hom g h p k w ll go o h mo l o ' hom l k o v o l v . f h mo l o urr l for g work h r g r r -of r o h hom g h p k w ll r v h mo l o ' hom g . h hom g h p ul h p k u l o h mo l o ' r -of r . po r v g h u l p k h mo l o w ll g up o h orr po o . h orr po o h h r of h p k r l o h mo l o u g rou g h r. g. 1.

Mo l v 2 . 10. orm g w r -of r

> “ f r g h h mov from o l k o o h r (. . urr f ul rou r h om u r h l h ov r w f ul rou r) mo l o L form w pr m r r -of r u g o of h o -l k u pr x v r h w

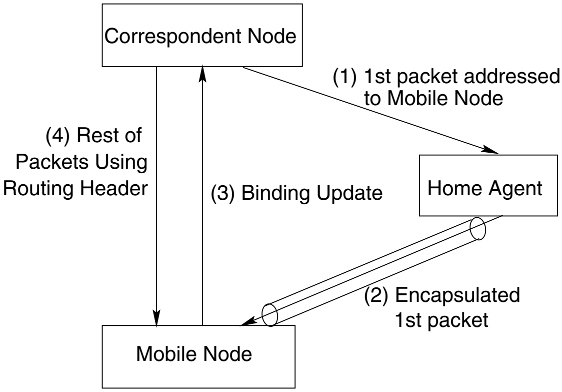

Correspondent Node
(1) 1st packet addressed to Mobile Node
Home Agent
(2) Encapsulated 1st packet
Mobile Node
(3) Binding Update
(4) Rest of Packets Using Routing Header

Fig. 1. ou g Mo l v

rou r. mo l o M form w pr m r r -of r
m x p h M o o oo fr qu l .”[1]

. 10. g g p o h om g

“ f r g o h g pr m r r -of r r
o 10. 10. mo l o M r g r h r -of r
w h hom g or r o m k h pr m r r -of r .”

o f w form r -of r u o m mo l o
l w hou r g r g o h hom g h h g l g lo
w ll r u . Mo l v l o p rm mo l o o form w pr m r
r -of r m o w m po po form g w pr m r r -of
r wh po l . ur ppro h o h o rv o w ll
r h x o .

3 Threshold-Based Registration (TBR)

o m mo l o l w hou form g w pr m r r -of
r h m mo l o mov o o h r u work w mplo h
m h m rou Mo l v o 10.9. l h g orw r g from
r v ou r -of r .

r w w rm olog *anchor home agent.* h mo l o
mov o for g l k r g r pr m r r -of r o hom
g h rou r h for g l k wh h h hom g of h
mo l o ’ pr m r r -of r w ll ll hor hom g . L r

[1] h k ywo “ L ” “ ” n “ ” qu ntly u n .
h y qu y o l v to nt p t n *Key words for use in RFCs to indicate requirement levels, RFC 2119* .

h mo l o l h *direct forwarding* from h hor hom g o urr r -of r . *immediate forwarding* m h p k r forw r from mo l o ’ pr v ou r -of r o h urr r -of r r Mo l v . 10.9.

h follow g u o w w ll ro u mplo m of h forw r g m h m g v g hr ppro h . h h r l propo l of h omprom of h r wo.

3.1 First Approach

h r ppro h mo l o l h forw r g o l from mm - pr v ou r -of r . r p u h forw r g u l h r h r o rou r h pr v ou v u h hom g or h um r of forw r g p x pr l m . h l m ll h m x mum um r of forw r g p F_{max}. h m h mo l o w ll form w pr m r r -of r u g h w r -of r g up o hom g . g. 2 for po l r o.

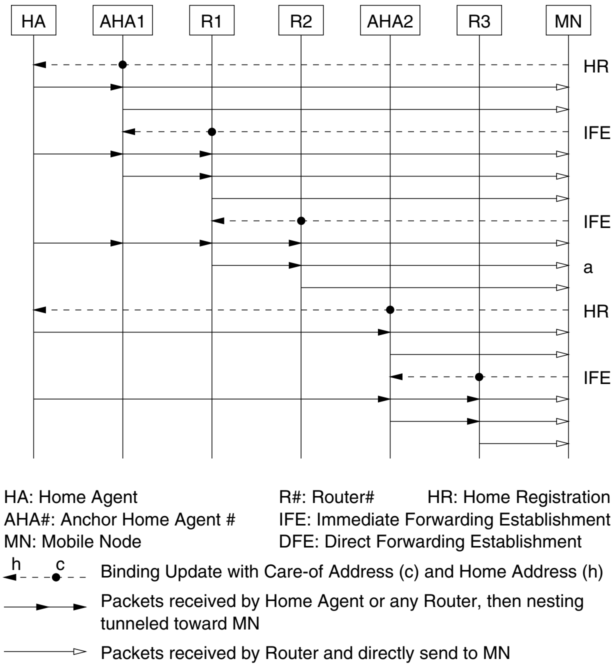

HA
AHA1
R1
R2
AHA2
R3
MN
HR
IFE
IFE
a
HR
IFE
HA: Home Agent
R#: Router#
HR: Home Registration
AHA#: Anchor Home Agent #
IFE: Immediate Forwarding Establishment
MN: Mobile Node
DFE: Direct Forwarding Establishment
h
c
Binding Update with Care-of Address (c) and Home Address (h)
Packets received by Home Agent or any Router, then nesting tunneled toward MN
Packets received by Router and directly send to MN

Fig. 2. l h forw r g from pr v ou r -of r o l . o h r po of r v p k *a* o l ho lo o mo l o ’ urr lo o r how h gur . h u o r g up o h orr po o wh h h p k o mo l o ’ mo r r -of r pr h ppro h ro u u l wh h m l o gm o of p k l for pro g. *Generic Packet Tunneling in IPv6 Specification* r omm h h f ul v lu of v u l p ul o L m . f p k w h o p o x o h r o g h l m op o w h v lu lo o orr po o h u l o mor h h um r of forw r g h p k w ll r for r h g h mo l o . hu h of h ppro h l m .

3.2 Second Approach

o h r ppro h woul h h mo l o l h forw r g o h from h mm pr v ou r -of r from urr pr m r r -of r wh mov o o h r u work. g. 3.

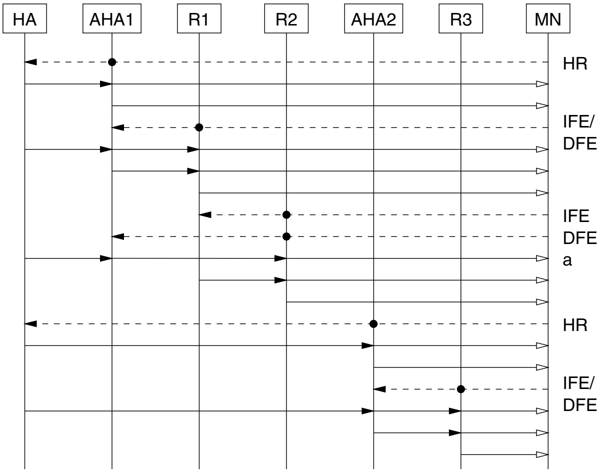


Fig. 3. l h forw r g from o h mm pr v ou r -of r h pr m r r -of r . *a* h p k u l from h hom g w ll o l o x r u l o h mo l o hu olv g h pro l m r g from h r ppro h. h r m ol r h m g. 2

o h h ppro h h l f m u h g up for l h g forw r g from mm pr v ou r -of r hor . h r o g r of r k g h h of forw r g h r lw r rou from h hor hom g o h mo l o . u h ppro h g r ou l g l g lo lo ll l o o um mor of h mo l o ’ r our h h r ppro h.

3.3 TBR — Final Proposal

hu w propo l ppro h h ppro h. h ppro h mo l o l h mm forw r g from h pr v ou r -of r h m mov o o h r u work. f r m ll um r of u ful mm forw r g p o IF_{max} h mo l o w ll l h r forw r g from urr pr m r r -of r . f r um r of u ful r forw r g p o DF_{max} h mo l o w ll r g r o hom g u g h l r -of r w pr m r r -of r . h mo l o w ll r g r h l r -of r o hom g l o wh v r h r m g l f m of h l hom r g r o r o h xp r r omm Mo l v . g. x mpl .

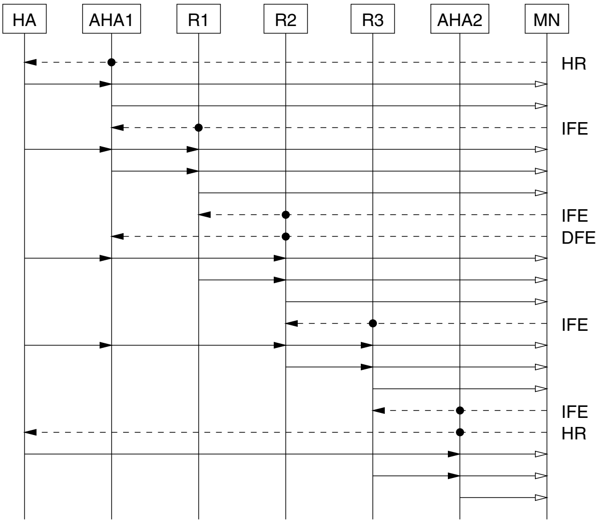


Fig. 4. Mo l o l h forw r g w h ppro h w h IF_{max} 2 DF_{max} 2. m ol r h m g. 2 g. 3

3.4 Tunneling Loop Avoidance

or r o vo p k u l loop w wo or v r l rou r mo l o houl r ur g hom r Mo l v o 10.1 wh w h o u work from wh h h forw r g ll v l . houl g up o h rou r w h z ro l f m o -r g r h pr v ou g for l h g forw r g from h mm pr v ou r -of r .

4 Discussion

h of h ppro h r follow

1. h v g l w op m z rou low g l g lo .
2. l l r l l . mo l o w ll h g hor hom g wh v r h l f m for hom r g r o or h l m um r of r forw r g p DF_{max} r h h lo for h hor hom g r u mo g h rou r h for g work. h h um r of mo l o h for g work r h r m ll r pro l h h hor f ul rou r w ll om o l k. h o of h rou r ou of or r h mo l o h g hor hom g up g pr m r r -of r g w r g r o o h hom g .
3. mpl m o . lmo ll h form o w o mpl m h ppro h lr or h g up l of h mo l o . l wo ou r (um r of mm forw r g um r of r forw r g) r o m k qu k o o wh h k of forw r g or r g r o houl p rform for h h o .

. r p r uppor . h mo l o mpl m h ppro h w h-ou pr - go o w h mo l o h for g work.

g up u for forw r g l hm houl u h h m w r for o h r g up h Mo l v . h ro u for mor ur k o m g o h h mo l o h hom g h for g work.

l hough w h v o om r h l l of h ppro h w h form l m ho u l rl horough p rform l l f for fu ur work or r o h op m l v lu for h p r m r IF_{max} DF_{max}.

5 Conclusion

h r l w pr h urr -of- h - r Mo l v o r -g o rou g p k forw r g. pr mo lgor hm for g up for r u g g l g lo wh l m g rou op -m l . h lgor hm mo l o l h mm forw r g from h pr v ou r -of r wh v r mov o o h r u work. f r v r x um r of mm forw r g p h mo l o l h r forw r g from pr m r r -of r . g f r v r x um-r of r forw r g p h mo l o r g r w pr m r r -of r o hom g . h u h ppro h h ov m o go l r h v og h r w h l l r l l mpl m o r -p r uppor h .

References

1. ng . n n . nt n t oto ol v on (v) p t on. 24 0. 1998
2. ohn on . k n . o l ty uppo t n v . nt n t t. o k n p og . 2000. http www. t .o g nt n t t t t mo l p pv 10.txt
3. k n . to o l ty uppo t. 2002. 199
4. t llu . ll L. ow n h l o l ty n g m nt m wo k. nt n t t. o k n p og . 1999. http www. t .o g nt n t t t llu uhmm m wo k 00.txt

. nl n o ng o ty th lo uly 11 uly 1 1999 o k g oup 2. .2 out ng o l o l o t (mo l p)

. n . K y wo o u n to n t qu m nt l v l . 2119. 199

. ont . ng . n k t unn l ng n v p t on. 24 3. 1998. 20

8. ynj¨l¨ . K . ng L. t on o h hol g t on lgo thm n o l v . u m tt to nt n t on l on n on two k 2000 (pt m 2000). 2000

Enhanced Mobile IP Protocol

Baher Esmat, Mikhail N. Mikhail, Amr El Kadi

Department of Computer Science, The American University in Cairo,
Cairo, Egypt
{besmat, mikhail, elkadi}@aucegypt.edu

Abstract. One of the most recent Internet challenges is to support transparent movement of people along with their computers, data and most of all applications. Therefore, Mobile IP has been developed to provide Internet mobility services.

This paper aims at enhancing the IETF Mobile IP standard. The model developed in this paper suggests a new caching mechanism, which is based on the Mobile Information Server (MIS). Actually, the MIS is designed to be part of the border router of any network that supports mobility services. Moreover, the paper suggests a *peering technique* by which information about mobiles hosts could be shared among different MISs. All the design issues including model components as well as mechanisms for caching and peering are described in details.

The simulation results show that the proposed design provides improved performance and better bandwidth utilization. The suggested architecture provides other qualitative advantages such as scalability and transparency.

1 Introduction

Mobile computing has assumed an increasing importance in recent years, and will pervade future distributed computing system. Although network standards were not designed with the capability of supporting the demand of mobility, the need is that they should grant the users a continuous access to their data, irrespective of their point of attachment. Mobile computing is still restricted by many obstacles [1].

As a mater of fact, the current IP version 4 [2] makes an implicit assumption that the point at which a computer is attached to the Internet is fixed, and its IP address identifies the network to which it belongs. The challenge is to develop a protocol, which allows computers to roam freely around the Internet and communicate with other stationary or mobile nodes, without major changes in the existing TCP/IP stack.

The mobility problem within the Internet is mainly concerned with the IP layer, since this layer handles all aspects related to addressing as well as routing. To illustrate this point [3], if a computer moves to another network, and retains its original IP address, this address will not reflect its new location, and consequently, all routed packets to this host will be lost. In the other hand if the mobile host gets a new address when migrating to another network, the IP address changes, the transport layer (i.e. TCP) connection identifier changes too [4], and hence all connections with this mobile host through its old address are going to be lost. Therefore, if the mobile

C.G. Omidyar (Ed.): MWCN 2000, LNCS 1818, pp. 158-173, 2000.

host moves without changing its address, it will lose routing, and if it gets new address, it will lose connections.

This paper is organized as follows. The next section presents the Mobile IP standard protocol. Section 3 describes the contribution of this work. An overview for the proposed design is going to be illustrated in section 4. In section 5, all the model components will be identified and discussed in detail. Next, section 6 presents all the simulation details as well as the results. Finally, section 7 concludes the paper.

2 Mobile IP

During the last few years, many contributions have been offered by different entities and groups, towards designing a model for a mobility supports Internet. The proposed models are different in terms of their components and methodology, but they are all aiming at keeping the mobile hosts communicating transparently via the Internet. Proposals from Columbia University [5,6], Sony [7,8], the Loose Source Routing(LSR) Proposal [9] as well as the Internet Engineering Task Force (IETF) Mobile IP working group [10,11,12], are the most outstanding models,

Since this work is an extension to the IETF Mobile IP standard, it is worth to focus on the operation of this standard protocol. First of all, it should be mentioned that the Mobile IP working group has been in charge of standardizing Mobile IP. Recently, the Mobile IP has become a standard, after passing through two stages [13]. The first one in which the base protocol was developed, with the objective that mobile nodes can roam transparently around the Internet, with no modifications whatsoever to other stationary nodes. The second phase has answered many open questions regarding the best route that the packet may take to reach a mobile node. This has been known by the *route optimization* problem.

2.1 Mobile IP Operation

According to [12], the IETF Mobile IP architecture defines special entities called the Home Agent (HA) and the Foreign Agent (FA), both cooperate to allow a Mobile Host (MH) to move without changing its IP address. Each MH is associated with a unique *home network* as indicated by its permanent IP address. Normal IP routing always delivers packets meant for the MH to this network. When an MH moves to a *foreign network*, the HA is responsible for intercepting and forwarding packets destined to the MH to anew address which is called the *care-of address*. The MH uses a special registration protocol to keep its HA informed with its new location.

Whenever a MH moves from its home network to a foreign network, or from one foreign network to another, it looks for a FA on the new network in order to obtain its new care-of address. In order for the MH to be able to work with this new address, it must go through a registration procedure via both, the foreign agent and the home agent. After a successful registration, packets arriving for the MH on its home network are *encapsulated* by its HA and forwarded to its FA. Encapsulation refers to the process of enclosing the original datagram as data inside another datagram with new IP header [14]. The source and destination addresses in the new header correspond to the HA and FA respectively. Upon receiving the encapsulated datagram, the FA strips off the new header and forwards the original one to the MH. This process at the FA end is known as *decapsulation*. If on the other hand, the mobile node needs to send a packet to any destination, the packet will be routed to its destination with the normal fashion without using either the home agent or the foreign agent. The figure below illustrates the operation of the mobile IP routing.

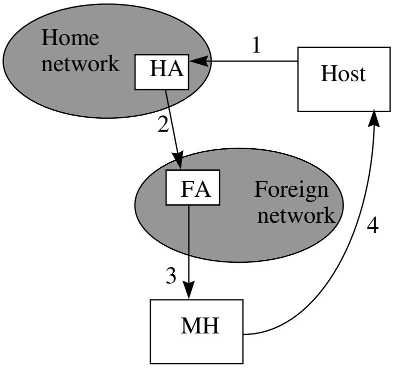


Fig. 1.

2.2 Route Optimization

As depicted in Figure 1, the IETF scheme has a major routing problem. Any packet which means to reach a mobile host, is directed first to the appropriate home agent, then to the foreign agent and finally received by the mobile host. This means that the above scheme does not allow the source host to reach the mobile host directly without passing by its home network. This problem is known as the *triangle routing problem*, and in order to solve it, a technique for route optimization is needed.

Route optimization [15] means solvingthe problem of triangle routing, by allowing for each host to maintain a binding cache for a mobile host wherever it is. When sending a packet to a mobile node, if the sender has a binding cache containing the care-of address of that mobile node, it will deliver the packet directly toward the mobile node, without the need to pass through the home network.

2.3 Mobile IP Problems

Although the IETF Mobile IP working group has enhanced its base protocol and provided a solution for route optimization, the protocol seems to have some deficiencies. From a performance point of view, the protocol intended to optimize the routing process, but realistically the routing has not been thoroughly optimized. For example, consider two computers connected to the same network, computer A and computer B. If the first one wants to reach a certain mobile host, it must go through the home network, at least for the first few packets, before reaching the mobile node.

Then, it caches the mobile node's address, so that it could reach it directly for the rest of the packets. Afterwards, if computer B needs to access the same mobile host, it must go through the same procedure again, as it does not have any cached information regarding the mobile host. The same process applies for any host on this network when trying to contact this specific mobile host. This implies that the first few packets directed from any host on a certain network toward a mobile host, are inefficiently routed through the home agent.

Another problem exist that relate to the fact that current implementation of Internet Protocol version 4 (IPv4) [2], which is currently running on all the Internet hosts worldwide, do not allow for any such mobile information to be cached. This means that in order for such a routing optimization to be achieved, every single host on the Internet must have its IP software modified.

2.4 Mobility Support in IPv6

IPv6 [16] has been developed with some sophisticated features that have not been supported by the current Internet Protocol (IPv4). IPv6 sustains major requirements concerning addressing, routing, security and mobility.

Mobility support in IPv6 [17] follows the same methodology that has been developed for IPv4. The same terminology is still valid as for home agents, mobile hosts, home and foreign networks, as well as encapsulation or tunneling. However, the term foreign agent is not of any more use. The reason is that any IPv6 is able to configure its own IP address automatically, as well as to choose its default gateway. This is accomplished via the Stateless Address Autoconfiguration [18] and the Neighbor Discovery protocols [19]. Therefore, it is quite straightforward that whenever an IPv6 mobile host migrates to any foreign network, it could easily detect the change in network connectivity, and configure its IP address automatically. Moreover, mobility within IPv6 borrows heavily from the route optimization specified for IPv4, which was described earlier in a previous section. By default, all IPv6 hosts are able to cache mobility information, via authenticated binding update messages. It is only the mobile host that has the authority to send binding updates to any other correspondent nodes.

Although people have been waiting for IPv6 to become the Internet standard [20], and many vendors have implemented IPv6 in their products for testing purposes, IPv6 is still under development. It is quite conceivable that the Mobile IP deployment will coincide with the standardization and implementation of IPv6 [21].

3 The Proposed Scheme

This work is intended to enhance the Mobile IP standard that has been developed by the IETF Mobile IP working group. Most of the Mobile IP protocol specifications are used in the development of this work.

This paper suggests a method for caching mobile information, different from that developed by the IETF working group. The proposed model implies suggests a *central cache engine* within each network, or a cluster of networks, responsible for caching mobile information. Moreover, all the functions performed by the HA's and

FA's, regarding encapsulation, decapsulation, registration and authentication, could be part of this central cache engine.

In addition, it is recommended for any network, or class of networks, connected to the Internet, to use its boarder router as a Mobile Information Server (MIS) which handles all caching as well as mobility services. As a matter of fact, designing a central caching mechanism does not necessarily imply that this should be part of the network router. Instead, building such a cache server, along with other mobility functions, can take place in any workstation in the network. However, this design is recommended for more than one reason. First, the proposed model manipulates the cached mobile information as part of the routing information that already exists in the routers, so that any cache entries are considered part of the routing table. . This new model allows for MISs to work in a peering fashion, by which mobility information can be exchanged. In addition, the model developed here aims at being transparent for the IP version used, whether it is IPv4 or IPv6. Eventually, the paper delivers a new caching mechanism, as part of the Mobile IP protocol. A complete practical architecture, with a simulation of all components and their functionality is delivered. Efficiency, scalability and transparency are the main value-added features in this new scheme, taking into account all security policies which have been addressed through the base Mobile IP model.

4 Design Overview

The proposed design is based on a centralized caching architecture. For a specific network, there is a cache server responsible for any mobile information concerning any node belonging to that network, or even it could cache other information regarding any external mobile node. In addition to caching, this server can handle all the functions of the home agent as well as the foreign agent, such as registration, authentication and tunneling procedures.

5 Model Components and Description

Figure 2 depicts the main components of the new suggested design. The figure illustrates four different networks (any networks that are members o the Internet for demonstration purpose) in order to describe the various functions and scenarios of this model.

5.1 Mobile Information Server (MIS)

The new model defines a new term called MIS. The MIS is suggested to be implemented in the border router. Border routers are basically responsible for routing the traffic between a group of networks and the outside world of the Internet. Moreover, In addition, border routers are now made responsible for other mobile services that were part of the home agent and the foreign agent in the Mobile IP scheme. Also, the new caching mechanism is designed to take place on these routers.

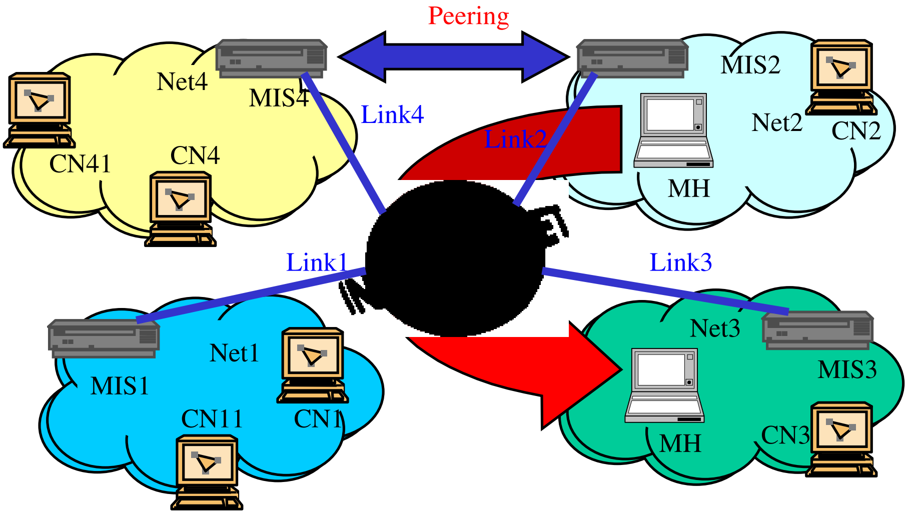


Fig. 2. New proposed model

Therefore, and because any of these routers handles a lot of other new services, it is going to be called in the context of this paper the MIS.

In addition to routing and caching functions, any MIS can be configured to work as a peer to another MIS. As a result of peering, MISs can exchange mobile information by the same concept of exchanging routing updates. Moreover, the MIS should include a table that encompasses all IP addresses of visiting mobile hosts, along with their Media Access Control (MAC) addresses, in order to deliver the packets to the proper destination after decapsulation. This table is called the *visitor list.*

5.2 Caching

It has been mentioned earlier that the MISs are responsible for caching mobile information. Actually, the cache entries are considered normal routing entries, with some extra fields for mobility.

Basically, the cache entry is suggested to include the following fields:

M O C P *old IP address* *next hop* *new IP address*

In this entry, the first four fields are flags which indicate the type of this routing entry. The following table describes briefly each of those flags.

The next two fields are similar to any ordinary routing information, they represent a destination IP address, as well as the next router in the way to reach this destination. In this case, the destination IP address is the old IP address of the mobile host. Finally, there is another field that indicates the new location of the mobile node.

Flag	Routing Entry
M	Set to 1 for all mobile routing entries
O	Set to 1 for mobile hosts controlled by the MIS
C	Set to 1 for mobile entries cached at the MIS as a result of cache update messages
P	Set to 1 for mobile entries cached at the MIS as a result of peering

Table 1. Mobile routing entry

5.3 MIS Peering

The MIS methodology provides another value-added service that is not in the Mobile IP protocol. This model allows for different MISs to share mobility information. In this case, the MISs seem to be working as peers, and the process of exchanging mobility information is called peering. As a matter of fact, the MIS in order to export any information regarding certain mobile host to another MIS, it should be the owner of this mobile host. In other words, the routing entry for this specific mobile node must have its *O* flag set to 1. When this routing entry gets to the peer MIS, it will have the *P* flag set to 1, in order to indicate that this information has come as a result of peering. Moreover, each MIS has in its Mobile IP configuration a list in which other MIS peers are defined. This means that MISs work as peers to each other based on some pre-defined routing configuration.

5.4 Mobility Scenarios

Reference to Figure 2, this section describes a scenario in which a mobile host *MH* migrates from its home network *Net2* to a foreign network *Net3*. The scenario shows how other correspondent nodes can reach *MH* while it is away from *Net2*, and even when it gets back.

1. *MH* migrates to *Net3* and gets a care-of address, which is actually the IP address of *MIS3*. Then, it sends a registration request to *MIS3*, which will relay this request to *MIS2*.
2. *MIS2* accepts the registration after checking the authentication extension included in the request. Hence, a registration reply is sent to *MIS3*, which in turn will inform *MH* with its status.
3. From now on and since the registration request has been accepted, *MH* becomes reachable via *MIS3*.
4. *CN1* wants to reach *MH*. So, it will start sending packets toward *Net2*. The packets arrive at *MIS2*, which could realize that they are destined to *MH*. *MIS2* contains a cache entry for the new location of *MH*. Actually, the *O* flag for this cache entry must be set to 1 because *MH* is owned by *MIS2*.
5. Right after the encapsulation, *MIS2* recognizes that *CN1* does not perceive that *MH* had moved. Therefore, *MIS2* sends *CN1* a cache update message. This cache update will be intercepted by *MIS1*, which in turn will update its routing table. In

this case, the cache entry at *MIS1* will have its *C* flag set to 1 since it has been generated through a cache update message.

6. *MIS3* receives encapsulated packets, and decapsulates them to *MH*, after obtaining its MAC address from the visitor list.
7. All packets from *MH* in their way back to *CN1*, are always routed directly to *Net1*, without necessarily passing by *Net2*.
8. *MH* decides to return to *Net2*. Consequently, it will ask for registration at *MIS2*.
9. If *MIS2* accepts this registration request, *MH* starts acting normally without any The next two fields are similar to any ordinary routing information, they represent a destination IP address, as well as the next router in the way to reach this destination. In this case, the destination IP address is the old IP address of the mobile host. Finally, there is another field that indicates the new location of the mobile node. mobility services.
10. *MIS2* sends a cache delete message to *MIS3* in order to release any routing information regarding *MH*.
11. *CN1* may try again to contact *MH*.
12. Packets are going to be tunneled through *MIS1* and directed to *MIS3*. From decapsulation, *MIS3* will discover that the destination address is not any more in its mobile visitor list.
13. *MIS3* sends a cache delete message to *MIS1*, which consequently is going to route the packets without any kind of encapsulation, through the ordinary route to *MIS2*.

In addition, Figure 2 depicts another correspondent node *CN11* that may need to access *MH* *net3*. Unlike the standard Mobile IP, all the packets from *CN11* toward *MH* will be routed directly to *MIS3*, since *MIS1* has a cache entry for *MH*.

Another difference between this new architecture and the Mobile IP one is the peering mechanism. Again from Figure 2, *MIS2* and *MIS4* work as peers to each other. Hence, *MIS4* will be notified that *MH* had moved. Therefore, it is possible that *CN4* can contact *MH* directly through a tunnel from *MIS4* to *MIS3*.

6 Simulation

Throughout the development of this work, simulation has been used to compare the Mobile IP standard protocol and the new proposed model. The main prominent difference between the two models is the caching methodology, by which the route optimization is verified, and the whole routing performance is induced. Actually, the influence of the caching mechanism can be evident in the total amount of traffic through the whole network, as well as the delay associated with packets while carried from source to destination nodes.

Therefore, this simulation has focused on the traffic as a main point for discrimination. All the quantitative results shown out of the two comparable models are in terms of packet delay as well as bandwidth consumption.

6.1 COMNET III

COMNET III [22] is a performance analysis tool which simulates computer and communication networks. It can be used to model both circuit switching and packet switching networks. In addition, it can accommodate different topologies of WANs and LANs in which many standards and protocols are supported, such as Ethernet, Token Ring, PPP, X.25, Frame Relay and ATM.

Regarding this work, COMNET III version 1.2 for Windows has been used to implement a number of models which differentiate between the standard Mobile IP architecture proposed by the IETF, and the model suggested in this paper. The simulated model presents all the issues described previously, concerning registration, tunneling, caching, route optimization as well as border routers which are known here by MISs. The networks described in the simulation were represented by Ethernet connections for LANs, Point-to-Point (PPP) links for WANs and processing nodes to simulate computers and workstations. All the networks are inter-connected using routers, in which user-defined routing tables are used to simulate the model. In addition, sending and receiving messages among the various nodes simulates the traffic. Finally, the model is verified and executed, and the results can be shown in graphs, or presented through reports of text format.

6.2 Network Components

This section provides a description for the network components simulated by COMNET III.

Processing node: All the simulated models use the processing node component to describe generic Internet hosts which are considered the endpoints of any Internet traffic.

Router node: Routers have been configured with *500 Mbps* bus rate, *50000 packet per second* as a processing rate, and the input and output delays are ignored. Moreover, all the models developed in this simulation have standardized on the static routing protocol, since using any dynamic protocols will make no difference to the results.

LAN connectivity: The IEEE 802.3 Ethernet standard is used.
WAN link: Point-to-Point Protocol (PPP) is used for all communication links with a bandwidth of *1.536 Mbps.*

Message source: Message sources are used to represent specific traffic based on the TCP/IP Protocol. A payload of *1460 bytes* and header of *40 bytes* is used for all the messages generated throughout the simulation. In addition, the message size may be changed based on the type of the message itself.

6.3 Simulated Models

This section describes a number of network architectures that have been simulated throughout this work. Generally, all the simulated models present the differences between the Mobile IP standard, and the new proposed scheme. The simulated models reflect the scenario that has been previously illustrated in Figure 2, in which a mobile host is migrating from one network to another whilst a correspondent node is trying to reach it. The models are simulated in simple as well as complicated structures. Simple models aim at presenting a preliminary overview for the Mobile IP operation, focusing on the main mechanisms and services for each architecture. On the other side, other more advanced designs are required in an attempt to simulate something close to reality. Such composite models include many nodes, routers and message sources that load the network with much more traffic.

6.3.1 Simple Models

This section illustrates the simplest cases for any Mobile IP architecture, where the simulated models consist of a single mobile node that migrates from its home network to another foreign one. Besides, there is a correspondent node belonging to another third network and it wants to get access to the mobile node. This scenario is shown for both the Mobile IP standard with route optimization support, and the new architecture developed within this paper. The most outstanding difference between the two schemes is the caching mechanism, as well as the fact that the border router within the new model is responsible for all mobile services.

Moreover, the simulation shows all the procedures defined by the Mobile IP standard. Such procedures include the registration request messages, registration acceptance, encapsulation as well as decapsulation. According to [3], the size of the registration message is *24 bytes* plus variable length extensions required for authentication. Likewise, the registration-reply is *16 bytes* beside those needed for authentication. As per our simulation, the registration messages are *48 bytes*, whereas the registration-reply messages are *40 bytes*, since extra bytes are used to indicate the variable length authentication extensions.

6.3.2 Composite Models

Similar to the simple models, in which the new proposed design has been compared to the Mobile IP standard, the composite models perform the same analogy accompanied by adding more components to the simulated models. In addition, the generated traffic is much more than that generated for the simple models.

Actually, the composite models contain five networks, two mobile hosts as well as many other correspondent nodes. Moreover, the new model simulates the peeing methodology that has been developed throughout this work.

6.4 Results

This section presents all the results that have been collected as an output from executing the simulated models. It will be noticed that all the models have been simulated over a simulation time of *60 seconds*.

As per the simple models, the point of discrimination between the two simulated schemes is the delay for the packets running from the source network to the foreign network where the mobile node is located. On the other hand, the evaluation of the composite models is based more the bandwidth consumption.

6.4.1 Simulation Results for Simple Models

As for the Mobile IP route optimization standard and as illustrated in Figure 2, when *CN1* talks to *MH*, the first few packets are going to be routed via *Net2*, then encapsulated toward *Net3*. The delay of the *CN1-Msg1* packets as well as that of the encapsulated packets *HA-Encap* are illustrated in Figure 3(a) and 3(b) respectively. The total average delay is *161.018 msec*.

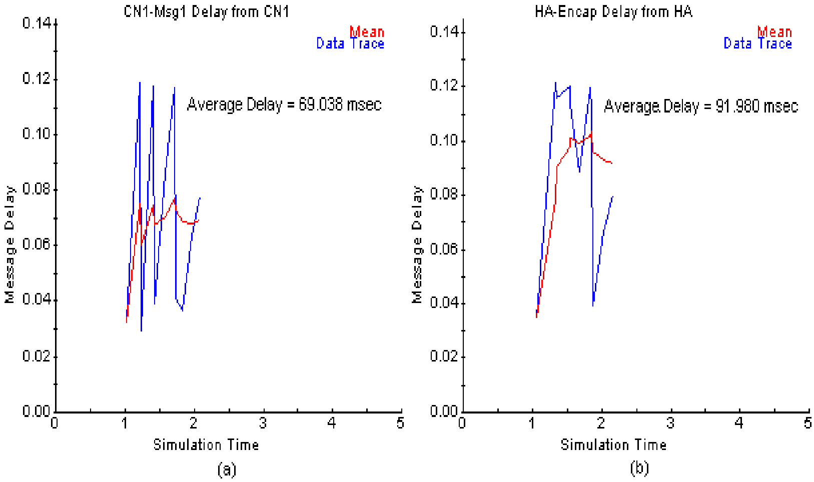


Fig. 3. Packet delay for *CN1-Msg1* and *HA-Encap*

But afterwards, *CN1* should get its cache updated and therefore packets are going to reach *MH* directly (CN1-Msg2) with an average packet delay of *94.544 msec*, as shown in Figure 4. The same scenario is applied for *CN11* when trying to access *MH*, causing almost the same results.

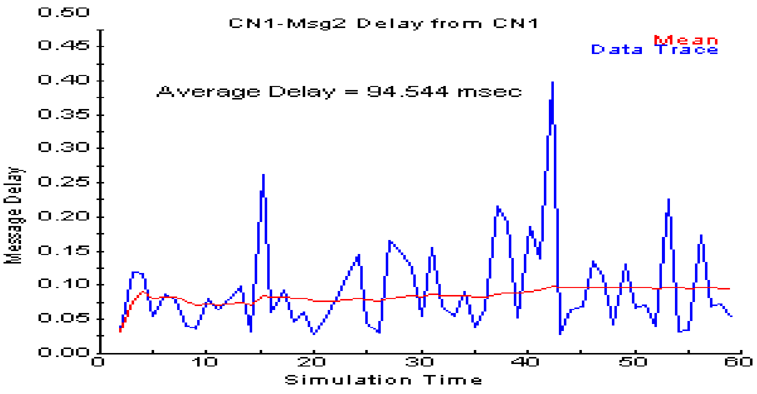


Fig. 4. Packet delay for *CN1-Msg1*

As formerly shown in Figure 2, *MIS1* is considered a central cache engine for *Net1*, responsible for any mobile information. Unlike the route optimization model, *CN11* may reach *MH* directly since the *MH* new care-of address is cached within *MIS1*. Therefore, the overall delay will be less than that of the last illustrated model. Figure 5 shows that in the new model, the average packet delay from *CN11* to *MH* is *79.681 msec*.

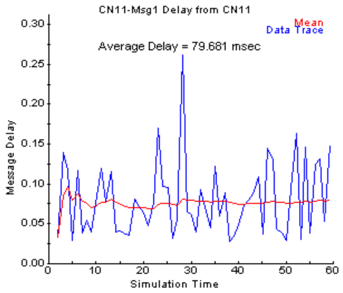


Fig. 5. Delay from *CN11* to *MH* in the new model

Although the difference in packet delays may be significant in such simple architectures, this might not be the case in real networks that carry millions of transferred packets per second. However, those models were basically simulated in order to prove that the proposed model has a quantitative advantage over the other models, even if this advantage is a minor issue in real applications. More importantly, this difference could be more significant from another perspective. For instance, if there is a certain application which ought the mobile node to stay in contact with a number of remote nodes at some other network, so that there are many nodes like *CN1* and *CN11* belong to the same network, and try to reach the same mobile host. In this case, it is better from a scalability point of view to have a central caching rather than storing the same information on many different machines.

6.4.2 Simulation Results for Composite Models

It has been stated before that for the composite models, the evaluation criteria is according to the bandwidth utilization. In fact, both composite models have been simulated with two networks working as home networks for two different mobile hosts, and each network has a single link to the Internet. The bandwidth utilization of each link is illustrated in this section.

Before presenting the results, it should be mentioned that COMNET III deals with any communication connection as a full-duplex link, in which the input bandwidth is independent of the output bandwidth. Therefore, it will be noticed that each link is represented by two graphs (a) and (b).

Assuming that the Internet link for the first home network is *Link A*, and for the other network is *Link B*. Figures 6 and 7 depict the channel utilization of *link A* in case the standard IP model and in the case of our new model respectively. Furthermore, Table 2 summarizes the results indicating that the new model is better than the standard one in bandwidth consumption.

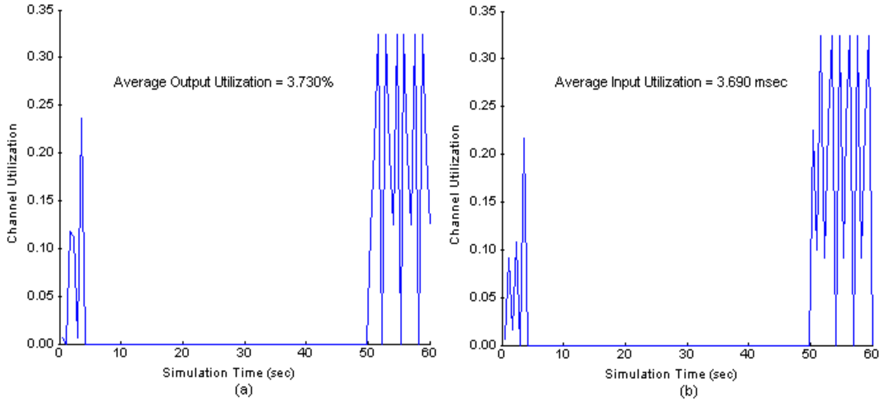


Fig. 6. Bandwidth utilization for *Link A* in the Mobile IP standard

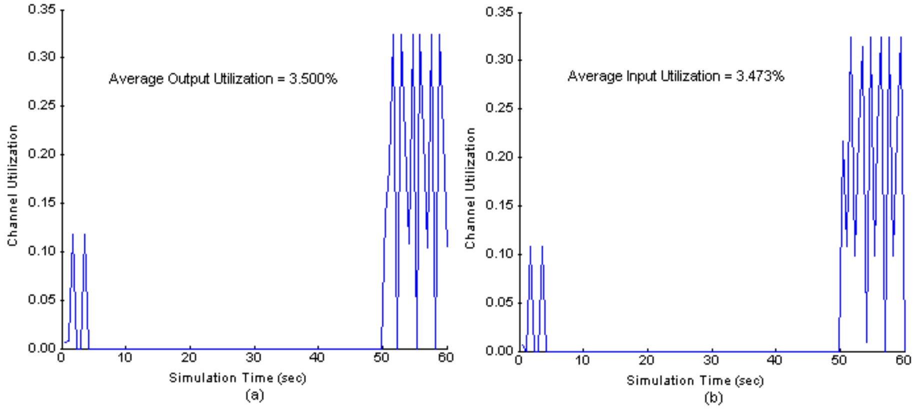


Fig. 7. Bandwidth utilization for *Link A* in the new model

Model	Link A		Total	Improvement
	In	Out		
Standard	3.690%	3.730%	7.420%	6.025%
New	3.473%	3.500%	6.973%	

Table 2. Improvement in bandwidth utilization for *link A*

As for the other home network connected via *Link B,* similar results are obtained and are represented in Figures 8 and 9 and summarized in Table 3.

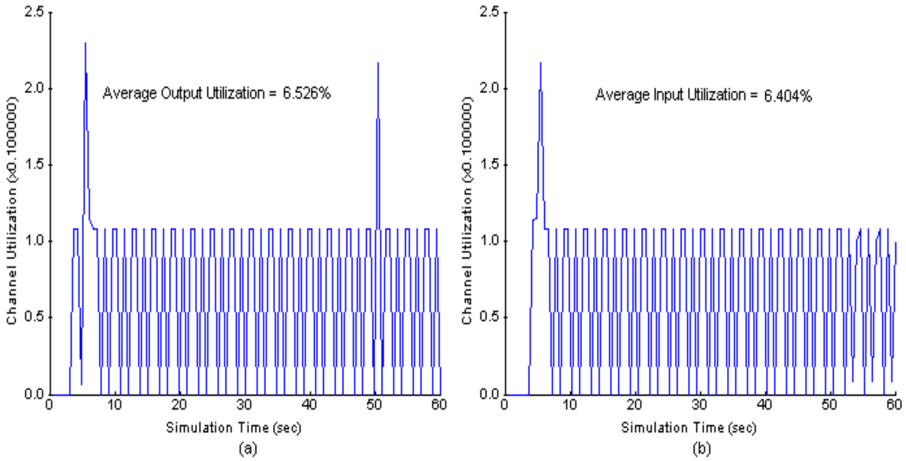


Fig. 8. Bandwidth utilization for *LinkB* in the Mobile IP standard

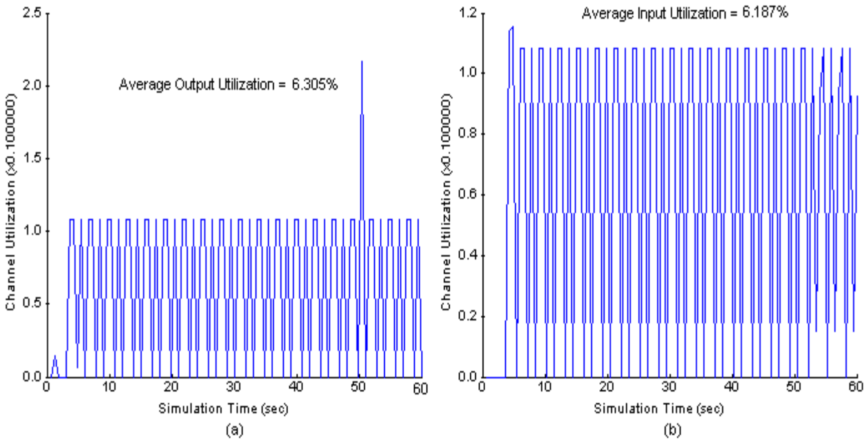


Fig. 9. Bandwidth utilization for *LinkB* in the new model

Model	LinkB		Total	Improvement
	In	**Out**		
Standard	6.404%	6.526%	12.930 %	3.387 %
New	6.187%	6.305%	12.492%	

Table 3. Improvement in bandwidth utilization for *link B*

Anyhow, such numbers prove the fact that the new model surpasses the standard one. Indeed the improvement ratio could be large or small depending on the number of the mobile hosts and the way they communicate with other nodes. But, the results prove that there is a quantitative improvement. Moreover, it should be mentioned that all the parameters that have been applied throughout the simulation are taken as an assumption, with the fact that changing such parameters will definitely change the output numerical results. However, any modification in the simulation parameters does not contradict with the fact that the new proposed model is quantitatively better than the Mobile IP standard.

7 Conclusion

In this work, we have suggested possible enhancement to the Mobile IP protocol that has been developed and standardized by the IETF. The IETF Mobile IP working group has proposed a technique for route optimization. Based on this concept, this paper has provided a new methodology for caching mobile information. Also, a new vital component called the MIS has been added to the mobile architecture.

Simulation results supports the logical expectation of improved efficiency when the new architecture is used over the standard one.

In addition to the quantitative gain, the new model has achieved other substantial qualitative advantages. From a scalability point of view, and after describing the details of the two mobile architectures, it is quite clear that in order to deploy such a wide area caching mechanism; some sort of centralized management is required, which is implemented in the MIS. Moreover, our new design is transparent not only to the Internet hosts, but also to the protocol used whether it is IPv4 or IPv6.

References

1. George H. Forman and John Zahorjan.: The Challenges of Mobile Computing., Computer Science & Engineering, University of Washington, 1993.
2. Postel J.B., Editor.: Internet Protocol. IETF RFC 791, September 1981b.
3. P. Bhagwat, C. Perkins, and S. K. Tripathi.: Network Layer Mobility: an Architecture and Survey. *IEEE Personal Comm.,* Vol.3, No.3, June 1996, pp. 54-64.
4. Postel J.B., Editor.: Transmission Control Protocol. IETF RFC 793, September 1981c.
5. J. Ioannidis and G. Maguire Jr.:The Design and Implementation of a Mobile Internetworking Architecture. *In Proceedings of Winter USENIX*, San Diego, CA, January 1993, pp. 491-502.
6. John Ioannidis, Dan Duchamp, and Gerald Q. Maguire Jr. : IP-based Protocols for Mobile Internetworking. Department of Computer Science, Columbia University.
7. F. Teraoka, Kim Claffy, and M. Tokoro.: Design, Implementation and Evaluation of Virtual Internet Protocol. *In Proceedings of the 12th International Conference on Distributed Computing Systems*, June 1992, pp. 170-177.
8. F. Teraoka and M. Tokoro.: Host Migration Transparency in IP Networks. *Computer Communication Review*, January 1993, pp. 45-65.
9. Y. Rekhter and C. Perkins.: Loose Source Routing for Mobile Hosts. Internet draft, July 1992.

10. Charles Perkins, Andrew Myles, and David B. Johnson.: IMHP: A Mobile Host Protocol for the Internet. *Computer Networks and ISDN Systems 27*, December 1994, pp. 479-491.
11. Charles Perkins and Andrew Myles.: Mobile IP. *SBT/IEEE International Telecommunications Symposium*, Rio De Janeiro, August 1994.
12. Charles Perkins.: IPv4 Mobility Support. IETF RFC 2002, October 1996.
13. Charles Perkins. Mobile IP: Design Principles and Practices. Addison-Wesley, 1997.
14. Douglas E. Comer.: Internetworking with TCP/IP Volume I: Principles, Protocols and Architecture. Prentice-Hall, 1995.
15. Charles Perkins and David B. Johnson.: Route Optimization in Mobile IP. Internet draft, ftp://ftp.ietf.org/internet-drafts/draft-ietf-mobileip-optim-07.txt, November 1997.
16. S. Deering and R. Hinden.: Internet Protocol, Version 6 (IPv6) Specification. IETF RFC (1883), December 1995.
17. Charles Perkins and David B. Johnson.: Mobility Support in IPv6. Internet draft, ftp://ftp.ietf.org/internet-drafts/draft-ietf-mobileip-ipv6-07.txt, November 1998.
18. S. Thomson and T. Narten.: IPv6 Stateless Address Autoconfiguration. IETF RFC 1971, August 1996.
19. T. Narten, E. Nordmark, and W. Simpson.: Neighbor Discovery for IP Version 6 (IPv6). IETF RFC 1970, August 1996.
20. Scott O. Bradner and Allison Mankin.: IPng Internet Protocol Next Generation. Addison-Wesley Publishing Company, 1995.
21. Charles Perkins.: Mobile Networking Through Mobile IP. *IEEE Internet Computing*, February 1998, pp. 58-69
22. COMNET III User's Mannual, Planning for Network Managers Release 1.2, 1996.
23 Mockapetris P.: Domain Names-Concepts and Facilities. IETF RFC 1034, November 1987.

A Reliable Subcasting Protocol for Wireless Environments

Djamel H. Sadok, Carlos de M. Cordeiro and Judith Kelner

Centro de Informática, Universidade Federal de Pernambuco, Caixa Postal: 7851
Cidade Universitária, Recife, PE, Brasil
{jamel, cmc, jk}@di.ufpe.br

Abstract. This paper presents an end-to-end reliable multicast protocol for use in environments with wireless access. It divides a multicast tree into sub-trees where subcasting within these smaller regions is applied using a tree of retransmission servers (RSs). RM2 is receiver oriented [1] in that the transmitter does not need to know its receivers, hence offering better scalability. The Internet Group Management Protocol (IGMP) is used manage group membership whereas the IETF's Mobile IP offers support to user mobility and a care-of address (temporary IP address). Each RS has a retransmission subcast address shared by its members and which may be dynamically configured using IETF's MADCAP (Multicast Address Dynamic Client Allocation Protocol) [8]. Most importantly, RM2 uses a dynamic retransmission strategy to switch between multicast and unicast retransmission modes according to the extra load generated in the network as well as the wireless interfaces by packet retransmissions. It is shown through both analytical modeling and simulation that RM2's dynamic adaptation is not only important but necessary when considering mobile access.

1 Introduction: Multicast in Mobile Environments

The IETF defines two approaches to multicast at the IP level, namely *bi-directional tunneling* and *remote subscription*. Other techniques for unreliable multicast have also been adopted in [4, 5, 6, 7]. The solution adopted in [4], and later on refined in [5], has scalability problems and assumes that group membership is static, which is hardly true when considering mobile environments. [6] and [7] deals with problems such as tunnel convergence, but does not deal with packet loss and performance issues.

Commonly used reliable protocols include the scalable reliable multicast (SRM) [2] and the reliable multicast transport protocol (RMTP) [3]. On the one hand, SRM is based on an application level framework where it is the application's responsibility to guarantee packet sequencing. This protocol is compared to RM2 using simulation later in the paper. On the other hand, RMTP defines a hierarchy of designated routers (DRs), a concept also used in RM2. Although each DR is responsible for handling error recovery within a region of the multicast tree, it does not say whether this is done in multicast or unicast and, therefore, does not concern itself with the control of retransmission overhead. This is a serious drawback when considering the emerging mobile environments.

C.G. Omidyar (Ed.): MWCN 2000, LNCS 1818, pp. 174-185, 2000.

In this paper, a new reliable multicast protocol, called RM2 (Reliable Mobile Multicast), tailored for mobile environments is presented.

2 Specification of the RM2 Protocol

The role of RM2 is to take a stream of packets generated by an application and deliver it to all mobile as well as fixed hosts interested in receiving it in a reliable and optimized way. Furthermore, RM2 guaranties sequential packet delivery with no packet loss to all its multicast members.

RM2 assumes that the network is formed of multicast routers and that cells are big enough to allow users to join and leave multicast groups. The RSs (Retransmission Servers) perform selective retransmissions on the basis of feedback in the form of negative acknowledgements they receive. RM2 dynamically establishes the subcasting regions while taking into account retransmission costs.

2.1 The Role of a Retransmission Server

A multicast sender must first divide a multicast message into smaller fixed size packets (except the last one). To each one of these packets, RM2 associates a sequence number ($\mathbf{n_{seq}}$). In order to guarantee end-to-end reliability, the receivers are required to send NACKs pointing out which packets they want to be retransmitted. In other words, RM2 implements selective packet retransmission. A NACK contains a sequence number N and bitmap B. N indicates that all packets with sequence number less than N have been correctly received by a given receiver. The bitmap B, on the other hand, shows which packets have actually been received. Consider the example where the ACK contains N=22 and B=01111101. In this case, the receiver wishes to indicate that it has correctly received all multicast packets with a sequence number less than 22 and that it is requesting the retransmission of packets 22 and 28 as indicated by the two 0s present in the bitmap B.

Sending NACKs to a multicast sender could lead to overwhelming it and causing an NACK implosion and even network congestion especially near the multicast source subnet. Therefore, RM2 divides the multicast network into hierarchical regions, where each one of these is controlled by a retransmission server (RS). This is responsible for gathering and processing NACKs from its region and for the retransmission of packets as requested by some receivers. Figs. 1(a) and 1(b) illustrate this concept. RM2 assigns RS functionality to fixed hosts selected on a network topology basis. The selection is subject to analysis in the following text.

The RSs are centric to RM2 support for reliable multicast sessions. Multicast packets are cached within the RSs buffers. The RSs are also responsible for the combination of NACKs received from lower RS hierarchies and hosts within their regions and responding to these when possible. If not, the retransmission request may be passed on to higher levels of RSs. Note that the separation between multicast routers and RSs is purposely done. It has the benefit of freeing router resources to the handling of multicast packets. RSs, on the hand, are better represented by hosts since they may need to use large buffering to keep copies of multicast packets for possible retransmission.

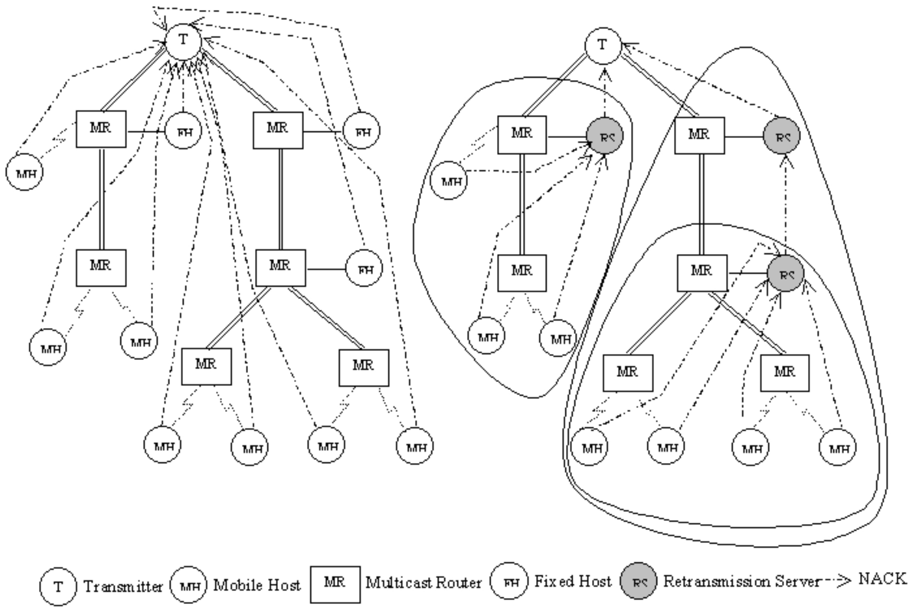


Fig. 1(a). Receivers sending NACKs to the multicast transmitter

Fig. 1(b). RS NACK processing

2.2 Establishing the RS Hierarchy

Core to the RM2 protocol is the establishment of a minimum cost spanning tree of RSs. The static selection of RSs in a WAN is on the basis of network topology. However, a multicast receiver selects dynamically its RS. Initially when there are no users in group a G, the RSs transmit a CFG_RS packet for G at each time interval T_{RS} in order to:

- Advertise its status as an RS to all potential members of a group;
- Define transmission costs where each CFG_RS packet contains a field with the cost of reaching the transmitter. This cost is null when the RS is not currently receiving data for G or that it is the actual transmitter. An example of a useful cost metric used in the simulation is the number of hops.

As soon as hosts join the multicast group they start receiving, in addition to multicast data packets, CFG_RS configuration packets. This way a host may be able to select the one that it is nearest to it on the basis on the cost information. The host then sends a REG_HOST to a suitable RS and sets a timer with a $2*T_{RS}$ value. The REG_HOST message has the following information:

- *HostType*: telling its RS whether it is a fixed or mobile host;
- *Group ID*: of the multicast group it is currently joining;

On receipt of a REG_HOST message a RS invokes the following actions:

- The RS updates its fixed/mobile host counters within its region and registers the hosts' unicast address;
- In the event that this is the first host to enter a multicast group at this RS, it dynamically allocates a D (multicast type) address, for example G1, using a protocol such as MADCAP [8]. This is the retransmission address for all receivers at this RS. Next the RS sends a RETR_ADDR packet telling its new host which multicast retransmission address its has to listen to. From now on, a new group G1 with a subcasting address has been established at this RS;
- Based on the number of retransmission requests a RS receives within a region, it may retransmit using either unicast or subcasting. This flexibility is a clear advantage over other protocols such as RMTP and SRM. Details of this mechanism are presented later in the paper. With the insertion of new hosts, new regions with separate RSs are formed. Each new RS performs the same procedure as a host in order to get a retransmission feed from a higher level RS. RM2 uses the Djikistra algorithm to establish the spanning RS tree where hop count is cost metric.

2.3 Support for Fixed Receivers

At the network level IGMP is used for group management. Existing multicast routing protocols such as the distance vector multicast routing protocol (DVMRP), protocol independent multicast (PIM) and multicast OSPF (MOSPF) may be used in conjunction with RM2 for routing multicast packets. A RS distinguishes between mobile and fixed hosts using the *HostType* field present in the REG_HOST registration message. Similarly, a host may leave a retransmission subcast group at a RS by sending it a LEAVE_GROUP message. The RS needs to keep track of the number of hosts in its region and should remove its link to the upper RS when it has no members to serve.

2.4 Support for Mobile Receivers

Mobile devices are somehow more delicate to handle at the RS. MHs send a normal IGMP join message to a multicast router (MR) using the care-of address from a foreign network (FN) obtained through the use of the IETF's mobile IP. The MR uses normal procedure to include itself in the IP multicast tree. The MH then sends a REG_HOST to its closest RS with the *HostType* field set to mobile. The following situations may happen as a result of host mobility:

- A MH sends an IGMP leave group message to its current MR before performing handoff. The MR may then check through a new IGMP query message to see if there are still multicast host members in its subtree. If not, it would remove itself from the multicast tree. Furthermore, the MH also sends a RM2 LEAVE_GROUP message to its respective RS in order to be removed from the retransmission subcast list;

- A second situation rises when the MH is unable to leave the group before handoff is complete or that the "leave" message is lost. Therefore, the MR may only know if there are hosts left at its subtree only after it sends a new IGMP membership query message. Since the IGMP leave message may itself be lost be lost, a registration timeout is associated to each MH. The RS relies on NACKs sent by MHs to know which of these remain active in its region. An interesting scenario emerges when a MH performs a handoff and moves to a new network (at another FA). This leads to packet loss due to the fact that the MH is unable to receive multicast packets when it is in between FAs. At reconnection, the MH must send an RM2 REQUEST message to start receiving from where it left the multicast transmission before the handoff.

2.5 Support for Mobile Transmitters

Reconfiguring the entire multicast tree each time a sending MH moves to a new FN is costly. Another problem results from the sending of NACKs to a FN where the MH no longer is. RM2 adopts two approaches to deal with these problems:

- When the MH is within its home network, it performs a link level multicast. The HA then forwards these downstream;
- If the MH is visiting a FN, it tunnels packets to its HA which then forwards these using local links and WAN interfaces. In other words, the HA in this case performs the role of a gateway. For performance and scalability reasons, RM2 limits the use of IP tunnels when there is a multicast mobile source transmitting.

2.6 RM2 Error Recovery

NACKs are used by hosts to signal lost and corrupt packets. RM2 adopts a similar approach to RMTP in that it collects NACKs per packet within a queue during a specific time interval and then retransmits the requested packets. This scheme clearly needs further tuning in the case of mobile hosts, especially considering that the requesting mobile hosts may have moved on to new FNs since they sent NACKs when dealing with high mobility users for example. RM2 adopts a dynamic retransmission technique at the RS level as shown in the following two scenarios:

- High error rate: representing environment with a high number of mobile users with wireless access;
- Low error rates: representing situations with a relatively low mobile to fixed users ratio;

A RS continuously monitors the number of fixed and mobile users and their NACKs. It is shown when describing the analytical model for RM2 that there are two main mathematical restrictions that control the adaptation of the retransmission algorithm that decides whether retransmitted packets at a RS are sent to all group members via multicast (more specifically subcast) or that they are sent via unicast to all those hosts that requested them. When the number of NACKs is relatively low, unicast is used at the RS to retransmit packets. RM2 monitors the network load, as shown in the analytical model, and should this becomes relatively high, it would then switch to multicasting retransmission packets. However, it is important that RM2 does not attempt to reduce network load at the detriment of low speed wireless interfaces. Indeed, RM2 only uses multicast as long as the wireless channel occupation remains bellow a threshold. Between overloading the fixed network and the wireless links, RM2 chooses to limit the traffic on the latter. There are, however, situations where, even when unicast is used at the RS, the wireless interface may see too many retransmission packets. The RS may then send, on behalf of its MHs, a REDUCE_FLOW message to the multicast source. Note that it is not mandatory for a transmitter to comply with this message. Table 1 shows a summary of RM2 messages.

Table 1. RM2 messages

Message	Originator	Receiver	Description
DATA	Transmitter	Receivers & RSs	Data Packet
REQUEST	Receivers & RSs	RSs	Requests the retransmission of lost packets
REPAIR	RSs	Receivers & RSs	Retransmission of lost packets
CFG_RS	Transmitter & RSs	Receivers & RSs	Building RS tree and establishing subcast regions
REG_HOST	Receivers and RSs	RSs	Host registration
RETR_ADDR	RSs	Receivers & RSs	Informs a region's retransmission subcast address
LEAVE_GROUP	Receivers & RSs	RSs	Indicates that a host is leaving a group or simply doing a handoff.
REDUCE_FLOW	RSs	RSs	Flow Control

3 Modeling Packet Retransmission

A *n*-ary network topology with a depth *h* is considered with equal fixed link costs and where retransmission requests are assumed to be uniformly distributed. If *PS* denotes the packet size, then equations (1) and (2) give the unicast and multicast retransmission costs to *K* receivers respectively.

$$C_U(K, h, PS) = K \text{ x } h\ x\ PS\ . \tag{1}$$

$$C_M\ (n, h, PS\ ) = n\frac{n^h - 1}{n-1}PS\ , n > 1\ . \tag{2}$$

The tree depth *h* depends is related to the total number of hosts (N_T) as shown in (3):

$$h(N_T, n) = \log_n\left(\frac{N_T(n-1)}{n} + 1\right)\ . \tag{3}$$

Define $P(Xf_{i,j})$ and $P(Xm_{i,j})$ as the probability that a fixed/mobile host *i* requests a retransmission of packet j. They are respectively given by:

$$P(Xf_{i,j}) = 1 - (1 - P(E_F))^h$$
$$P(Xm_{i,j}) = 1 - (1 - P(E_F))^{h-1}\ (1 - P(E_M))\ \ .$$

where E_F and E_M are packet error rates for fixed and wireless links.

Let $P(Xf_{i,j})$ and $P(Xm_{i,j})$ be the probabilities that a fixed or mobile host i requests the retransmission of packet j respectively. Furthermore, let $P_i(Y = K_F, Z = K_M)$ represent the probability of having exactly K_F and K_M retransmission requests of a given packet i for both fixed and mobile hosts, while N_F is the total number of fixed hosts and N_M the total number of mobile hosts. Since retransmission request events are independent and that $P(Xf_{i,j})$ and $P(Xm_{i,j})$ remain practically constant during the experiment, $P_i(Y, Z)$ may be modeled as a binomial distribution. Therefore, we have:

$$P_i(Y = K_f) = \begin{pmatrix} N_F \\ K_f \end{pmatrix} P(Xf_{i,j})^{K_f}(1 - P(Xf_{i,j}))^{N_F - K_f}\ \ . \tag{4}$$

for fixed hosts, and

$$P_i(Z = K_m) = \begin{pmatrix} N_M \\ K_m \end{pmatrix} P(Xm_{i,j})^{K_m}(1 - P(Xm_{i,j}))^{N_M - K_m}\ \ . \tag{5}$$

for mobile hosts.

Through consecutive mathematical refinements of equations (4) and (5), we obtain equations (6) and (7) as representing packet retransmission in both unicast and multicast modes:

$$\begin{aligned} E(CR_U) = hPS(N_F P(Xf_{i,j})(1 - P(Xf_{i,j})) + (N_F P(Xf_{i,j}))^2 + \\ N_M P(Xm_{i,j})(1 - P(Xm_{i,j})) + (N_M P(Xm_{i,j}))^2)\ \ . \end{aligned} \tag{6}$$

$$E(CR_M) = n\frac{n^h - 1}{n-1}PS(N_F P(Xf_{i,j}) + N_M P(Xm_{i,j}))\ \ . \tag{7}$$

Fig. 2 presents the network load behavior as the number of retransmission requests arriving at the transmitter increases (equations (i) and (ii)). We see that, as the number of requests grows, the overload generated by the retransmitted packets depends on the retransmission mode (unicast or multicast). Ideally, RM2 should change into multicast retransmission when the two curves intersect as shown in fig. 2. Table 2 shows some of parameters used for this analysis. Through the variation of the ratio mobile/fixed hosts, we show that the effect of mobile hosts is far more important on retransmission load and that the initial unicast retransmission scheme cannot be maintained when there is a large network.

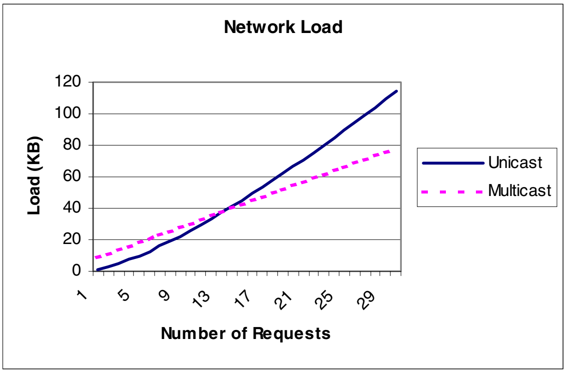


Fig. 2. Network Load

Table 2. Some Parameters for the Analytical Evaluation

Parameter	Value
$P(E_F)$	10^{-9}
$P(E_M)$	10^{-3}
n-arity	2
PS (Packet Size)	1KB

Actually, the RM2 retransmission mechanism takes into account two restrictions: R1 and R2. R1 guaranties that the network load does not exceed an established threshold *p* whereas R2 ensures that wireless retransmission channel utilization is not overwhelmed by multicast retransmitted packets. R1 is given below (equation 8):

$$\alpha.E(CR_U) \leq p \ . \qquad \textbf{(8)}$$

Where α:

$$\alpha = \frac{1}{N_M} . \frac{N_M P(E_M)}{N_M P(E_M) + N_F P(E_F)} = \frac{P(E_M)}{N_M P(E_M) + N_F P(E_F)} \ . \qquad \textbf{(9)}$$

In other words, α represents the average error rate related to fixed and wireless errors as well as the number of both user types. Fig. 3 shows the impact of both fixed (N_F) and mobile users (N_M) (equations (vi) and (vii)). Through fig. 3, we see that mobile users have a much bigger effect on the load generated by duplicate packets than fixed users. Therefore, RM2's retransmission mechanism was specially adapted to deal with mobile users.

R2, on the other hand, is given by:

$$E(CR_M) \leq E(CR_U)\ . \tag{10}$$

That is to say that RM2 must always check if it is not time to switch from a unicast to multicast retransmission in order to lower the network packet retransmission load.

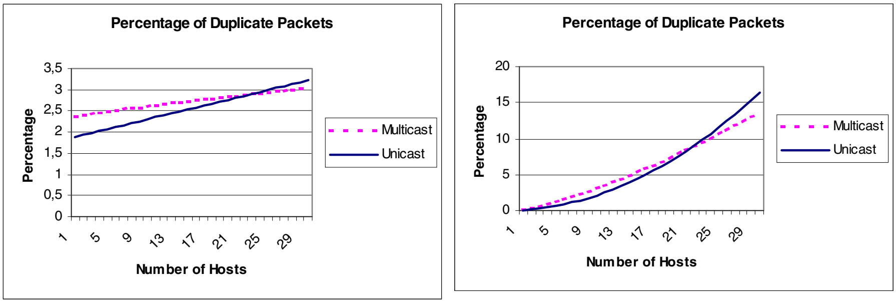


Fig. 3. Impact of Fixed and Mobile hosts

4 RM2 Simulation

RM2 was implemented in the U.C. Berkeley/LBNL ns-2 simulator. The source code is available from *http://www.di.ufpe.br/~cmc/research/rm2/rm2-source.zip*.

Topology. A backbone topology with 7 multicast routers has been used with 2 Mbps links and 10ms delay. The scenario consists of a maximum of 50 users joining a multicast transmission at different levels of the hierarchy. Fixed access is characterized by a 10Mbps speed and 5ms delay, whereas mobile access uses 14 kbps and a 50ms delay. We used ns-2 DVMRP (Distance Vector Multicast Routing Protocol) for building routing tables, IGMP for group management and mobile IP for address allocation.

Transmission. Each simulation experiment starts at time 1 second, and after the bootstrap phase of 8 seconds a CBR (constant bit rate) source starts transmitting 5000 1KB packets at 64 Kbps.

Error Model. In RM2's simulation, errors are a result of buffer overflow in routers and transmission packet error rates. Whereas the first one represents the dominant source for Internet packet loss, the second one reflects the error probability inherent to each link, as explained in the analytical model and configured in table 2. As presented earlier, the analytical model determines that the loss rates receivers experience are obtained by compounding the loss rates on the links from the sender to the receivers.

Results. The results mainly validate the analytical modeling earlier shown. The network load shown in fig. 4 presents similar tendencies to the one resulting from the analytical model presented fig. 2.

Similarly, the results show that the impact (see fig. 5) of both fixed and mobile users is in line with the one from the analytical model shown in fig. 3.

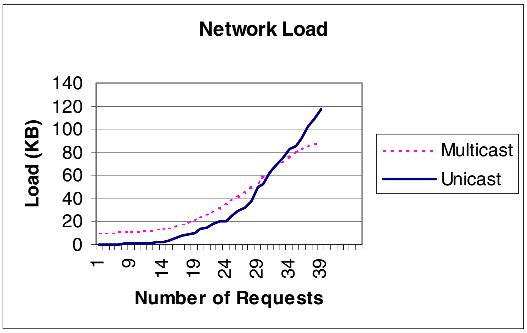


Fig. 4. Simulation of Network Load

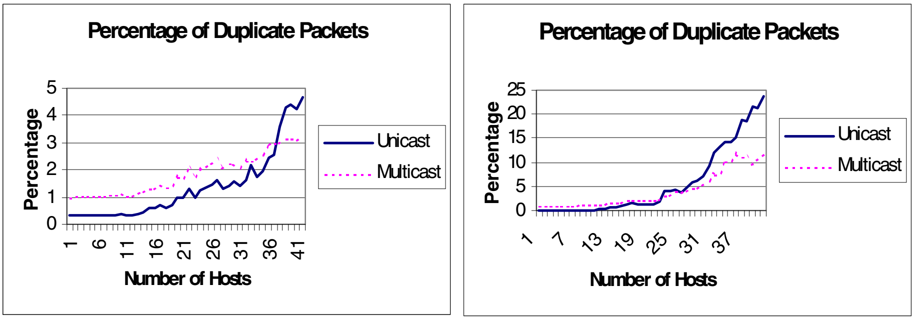


Fig. 5. Impact of Fixed and Mobile Hosts

5 Comparing RM2 Performance with SRM

Finally, we compare RM2 and SRM performance under similar loads. SRM has been selected since it is well referenced in the literature and readily available in the ns-2 simulator. We set R1 as a condition for RM2 to ensure that wireless interface may not see more than 20% of retransmission. Fig. 6 illustrates link utilization during the simulation.

The results are somehow intuitive, SRM generates higher levels of network occupancy since it always retransmits in multicast mode which may of course overload low speed wireless interfaces (in fig. 6, there are times where SRM uses almost 80% of this capacity).

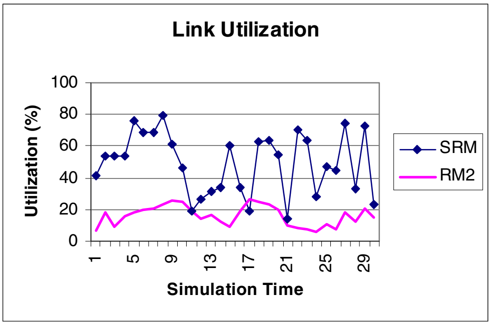


Fig. 6. Utilization of Wireless Links in SRM and RM2

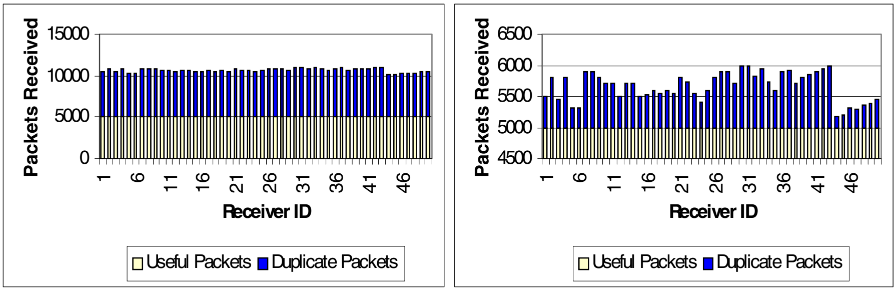


Fig. 7. SRM Duplicate Packets (left) and RM2 Duplicate Packets (right)

Furthermore, SRM's average occupation of wireless interfaces was 50% compared to a mere 16% when using RM2. Fig. 7 shows the number of duplicates seen by a host. Since SRM uses multicast for retransmitting packets to all group members, it presents a large overhead (close to 100%). This is often unacceptable for mobile hosts specially when compared to the overhead of RM2 which is less than 20% (maintaining restriction R1). Clearly, RM2 retransmission adaptation not only optimizes network traffic load but more importantly spares low speed wireless links.

6 Conclusions

This work has presented a new reliable multicast protocol, suitable for use in wireless environments. RM2 defines a hierarchy of retransmission servers which implement subcasting. It was shown through analytical modeling and simulation that the adaptation mechanism reduces multicast traffic due to packet retransmission and saves link utilization at the wireless interface. Finally, through simulation, RM2 efficiency over SRM was shown to be superior.

References

1. S. Pingali, D. Towsley, J. Kurose, “A Comparison of Sender-Initiated and Receiver-Initiated Reliable Multicast Protocols”, *Proc. ACM SIGMETRICS Conf. On Measurement and Modeling of Computer Systems*, May 1994
2. S. Floyd, V. Jacobson, S. McCanne, C. Liu, L. Zhang, “A Reliable Multicast Framework for Light-Weight Sessions and Application Level Framing”, *ACM SIGCOMM’95*, *Conf. on Applications, Technologies, Architectures and Protocols for Computer Communications*, August 1995
3. S. Paul, K. Sabnani, J. Lin, S. Bhattacharrya, "Reliable Multicast Transport Protocol (RMTP)", IEEE Journal on Selected Areas in Communications, April 1997
4. A. Acharya, B. Badrinath, “Delivering Multicast Messages in Networks with Mobile Hosts”, *Proc. Of the 13th International Conference on Distributed Computer Systems*, May 1993
5. A. Acharya, B. Badrinath, “A Framework for the Delivery of Multicast Messages in Networks with Mobile Hosts”, *Wireless Networks*, 1996
6. V. Chikarmane, C. Williamson, R. Bunt. W.Mackrell, “Multicast Support for Mobile Hosts Using Mobile IP: Design Issues and Proposed Architecture”, *ACM/Baltzer Mobile Networking and Applications*, 1997
7. T. Harrison, C. Williamson, R. Bunt. W.Mackrell, “Mobile Multicast (MoM) Protocol: Multicast Support for Mobile Hosts”, *Department of Computer Science*, University of Saskatchewan, Canada
8. MADCAP Protocol, http://www.ietf.org/internet-drafts/draft-ietf-malloc-madcap-07.txt

Author Index